Phenotypic Variation and Robustness in Complex Genetic Systems

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde

(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Tugce Bilgin Sonay

aus der Türkei

Promotionskomitee

Prof. Dr. Andreas Wagner (Vorsitz)

Prof. Dr. Mark Robinson

Dr. Michael Krützen

Zürich, 2014

Abstract

Phenotypic variation with a genetic basis is the raw material for natural selection and evolution. While it is created by genotypic change, phenotypes are also to some extent robust to genotypic perturbations. Such robustness exists on multiple levels of biological organization. To understand the origins of phenotypic variation with a basis in genetic change, one needs to understand how genotypic changes map to phenotypic changes. In this dissertation I study phenotypic variation and robustness on various levels of biological organizations. I ask how genotypic properties of synthetic metabolisms, such as network size and number of utilizable carbon sources map to metabolic phenotypes, that is, biomass synthesis rates. I describe the trade-offs between these properties quantitatively and show that they can explain most of the variation in synthesis rates of a metabolism. The observations I make are also relevant for synthetic metabolism design, which aims at large-scale, fast, and efficient synthesis of pharmaceuticals, chemical reagents, and biofuels. In a second project, I ask to what extent physicochemical changes in amino acid properties or in protein folding caused by mistranslation affect the codon choice of organisms. I find evidence that selection has increased the incidence of robust codons for ligand-binding amino acids, which suggests that it can affect the robustness of very small units of biological organization. A third project focuses on how tandem repeat instability relates to gene expression divergence in primates. I observe that genes with tandem repeats in gene regulatory regions are associated with high expression divergence. Hence, tandem repeats may contribute substantially to gene expression evolution in primates. Since tandem repeat instability is a hallmark of colorectal tumors, the final project compares phenotypic consequences of tandem repeat instability in gene promoters of tumor and

normal tissues. Repeat instability is enhanced in tumors compared to healthy tissues. Those genes with repeat instability are significantly overexpressed. These findings suggest an important role for tandem repeat instability in the differential gene expression observed for colorectal tumors.

Zusammenfassung der Dissertation

Phänotypische Variation verursacht durch genetische Variation liefert die Grundlage für die natürliche Selektion und die Evolution. Und obwohl genotypische Veränderungen zu neuen Phänotypen führen können, sind bestehende Phänotypen in gewissem Umfang auch robust gegenüber solchen genotypischen Störungen. Diese Robustheit zeigt sich auf verschiedenen Ebenen der biologischen Organisation eines Organismus. Um den Zusammenhang zwischen phänotypischer und genetischer Variation zu begreifen, muss man insbesondere verstehen, wie genotypische Veränderungen sich auf den Phänotyp auswirken. In dieser Dissertation untersuche ich die phänotypische Variation und Robustheit auf verschiedenen Ebenen der biologischen Organisation eines Organismus. Ich frage danach, wie genotypische Eigenschaften von synthetischen Stoffwechseln, wie zum Beispiel Netzwerkgröße und Anzahl nutzbarer Kohlenstoffquellen, sich auf metabolische Phänotypen auswirken, beispielsweise auf die die Syntheseraten von Biomasse. Ich beschreibe die Kompromisse zwischen diesen sich teilweise widersprechenden genotypischen Eigenschaften, und zeige, dass dadurch der grösste Teil der Variation von Syntheseraten des Stoffwechsels erklärt werden kann. Die Beobachtungen, die ich mache, sind auch für das Design künstlicher Stoffwechsel von Belang, wo es um die schnelle und verlässliche Synthese im grossen Umfang von Pharmazeutika, chemischen Reagenzien, und Biokraftstoffen geht. In einem zweiten Projekt untersuche ich, inwieweit physikalisch-chemische Veränderungen der Eigenschaften von Aminosäuren oder Übersetzungsfehler bei der Proteinfaltung die Auswahl von Codons beeinflussen. Ich finde Hinweise darauf, dass durch natürliche Selektion vermehrt robuste Codons für Aminosäuren verwendet werden, welche Liganden binden. Dies bedeutet, dass natürliche Selektion die Stabilität sehr kleiner Einheiten

der biologischen Organisation erhöhen kann. In einem dritten Projekt konzentriere ich mich auf die Instabilität von Tandem-Repeats und untersuche deren Zusammenhang mit den Unterschieden in der Genexpression bei Primaten. Ich zeige, dass Gene mit Tandem-Repeats in genregulatorischen Regionen mit grossen Unterschieden in der Genexpression verbunden sind. Tandem-Repeats könnten also zur Evolution der Genexpression bei Primaten einen wesentlichen Beitrag leisten. Da die Instabilität von Tandem-Repeats ein Kennzeichen kolorektaler Tumore ist, vergleiche ich im letzten Projekt phänotypische Folgen dieser Instabilität in Gen-Promotoren von Tumorgeweben und von gesunden Geweben. Die Instabilität von Tandem-Repeats ist in Tumorgeweben erhöht. Jene Gene, deren Instabilität in den Tandem-Repeats deutlich grösser ist werden überexprimiert. Diese Ergebnisse deuten darauf hin, dass die Instabilität von Tandem-Repeats eine wichtige Rolle in der Differenzierung der Genexpression spielt, welche in kolorektalen Tumoren beobachtet werden kann.

Acknowledgments

Throughout these four years, I learned many things, most important ones being not really scientific. I went to countless countries, met wonderful people, got my first pet, dyed my hair to red, saw the ocean for the first time, survived a risky operation, started writing, had the happiest moments, appreciated friendship truly, and got married. In short, I became more of myself. Numerous people contributed to this path, to whom I owe many thanks.

I am sincerely grateful to my supervisor Andreas for giving me the opportunity to work in his lab, and for the scientific freedom he gave me. In a way, he's the one who gave start to this path and supported me throughout. I thank my ex-supervisor Işıl Aksan Kurnaz for her encouragement, and faith in me and also for her endless support. Special thanks to Maria Anisimova for her valuable guidance and sincere companion. I am grateful to my co-advisor Mark Robinson, who was always there, whenever I was lost in the statistics. I owe a big thanks to my co-advisor Michael Krützen. He not only gave me some brilliant ideas on the primate project but also thanks to him, I know my next job, which gave me strength and motivation to finish my thesis.

I thank many members of the Wagner lab, especially to Joao, Evandro, Aditya, Niv and Adrian. They helped me countlessly in my projects and had always time for a scientific or nonscientific chat, whenever I needed. Special thanks to Elina for being the companion I always wanted to have. She brought sunshine to my office with her incredible personality. I am grateful to Mehmet Somel and Murat Tugrul, my

conference buddies, who made conferences so fun and taught me a lot of things at the same time.

I gratefully thank all my friends, especially to Margot, Deniz, Akos, Ale, Natasha, Valentina, Mariana, Macarena, Marta, Elena and Kasia who helped me finding my way numerous times. I am deeply grateful to my parents for their endless love and encouragement. They made me who I am today. Finally, I am indebted to my best friend and love Ali for his kindness, unlimited patience and wise directions. He gave meaning to my life.

This thesis is dedicated to all these individuals, who crossed my life and changed me forever. Thank you.

Table of Contents

1. General Introduction

The genotype holds an organism's full complement of hereditary information (Tautz and Schmid 1998), whereas the phenotype comprises its traits. Metabolic activities, gene expression, protein folding, morphology, and behavior are examples of phenotypes. The distinction between genotype and phenotype is fundamental in evolutionary biology. The phenotype determines an organism's chances of survival and reproductive output; it is the raw material for natural selection and evolution (Bull 1987; Mitchell-Olds et al. 2007; Stranger et al. 2007a). The inheritance of the phenotype occurs mostly as a secondary consequence of the inheritance of genetic material (Benfey and Mitchell-Olds 2008; Frazer et al. 2009; Lehner 2013).

Phenotypes of biological systems are to some extent robust to genotypic changes, that is they remain unchanged after a perturbation. Such robustness exists on multiple levels of biological organization. Therefore, to understand the origins of phenotypic variation with a basis in genetic change, one needs to understand first the fundamental properties of robustness, and then explore how genotypic changes map to phenotypic changes.

1.1. Robustness

Biological systems are continually subject to mutation and environmental variation. A biological system is robust if it continues to function in the face of these perturbations (Kitano 2004). Proteins can tolerate many amino acid changes; metabolic pathways continue to sustain life even after removal of important enzymes; drastic changes in

embryonic development can lead to an essentially unchanged adult organisms (Kitano 2004; Wagner 2005a). The mechanisms underlying robustness are diverse, ranging from thermodynamic stability at the protein level to behavior at the organismal level. I will give some examples of robust systems in various contexts, and then describe evolutionary origins, and mechanistic causes of robustness. I will also list some approaches to study robustness, particularly metabolic models, as the analyses in Chapter 2 are based on this approach. I will also describe various strategies biological systems apply for achieving a robust translational machinery, because Chapter 3 of this dissertation deals with selection for robust translation.

1.1.1. Examples of robustness to various perturbations

Robustness to genetic perturbations

A key example of robustness to genetic perturbations can be found in proteins. Proteins can be quite tolerant of genetic mutations. Through selection this can lead to highly diverse sequences that fold into similar structures and perform conserved biochemical functions. For example, Huang and collaborators (Huang et al. 1996) showed that point mutations in 84 percent of the amino acids of an *E. coli* beta lactamase do not have any severe effect on the protein function. In human 3-methyladenine DNA glycosylase, 66 per cent of single amino acid substitutions do not disrupt function (Guo et al. 2004). Even in the highly conserved catalytic core regions of proteins, approximately one-third of amino acid sites can tolerate substitutions (Guo et al. 2004; Materon and Palzkill 2001).

Robustness to non-genetic perturbations

Non-genetic perturbations comprise two components. The first one is noise, random variability of quantities important to cell functions (Pilpel 2011). For example, cells that are genetically identical, may occur within the same tissue can have different expression levels of proteins, different sizes, and structures (Ladbury and Arold 2012; Stewart-Ornstein et al. 2012). Noise, in general, can be an obstacle in tuning a system to the "fittest" state and maintaining it there (Pilpel 2011). Therefore, a phenotypic trait that is associated with fitness is expected to be to some extent robust against such stochasticity. For instance, developmental gene expression is extremely similar in a given cell type from one individual to another. Examples include *Hox* genes, which are responsible for anterio-posterior positioning that determines which cells will form which body structures, such as legs or antennas in the fruit fly Drosophila (Pearson et al. 2005b). Several studies demonstrated robust expression in essential genes compared to non-essential ones (Newman et al. 2006), in dosage-sensitive genes (Batada and Hurst 2007), and in genes that produce a strong growth defect when deleted compared to those producing a weak growth defect (Batada and Hurst 2007).

The second component of non-genetic perturbations is environmental change, including change in temperature, nutrient, oxygen supply, water availability and soil conditions. A recent study (Oliveira et al. 2014) revealed a striking example of robustness against environmental perturbations in Drosophila development. Oliveira and colleagues showed that expression of tissue patterning genes always aligned with certain milestones of development, such as moulting and pupariation even if developmental times were altered through temperature or hormone synthesis. Another example involves microRNA, small non-coding RNA's which function in gene

expression regulation. A highly conserved microRNA, miR-7, functions in several feedback and feedforward loops of animal developmental regulatory networks to buffer them against perturbations (Li et al. 2009b).

1.1.2. Evolutionary scenarios for the origins of robustness

The adaptationist scenario

Robustness can have three evolutionary origins. One of them is that robustness can be an adaptation. Such adaptive robustness evolved primarily to ameliorate the detrimental effects of genetic mutations, of environmental change, or of both (Fisher 1928). Because almost all phenotypic variation represents a deviation from the optimum for a well-adapted trait, any mechanism that limits phenotypic variation should be favored by natural selection. A compelling example comes from adaptation of wild yeast strains to oxygen-limited environments (Fidalgo et al. 2006). The yeast gene *FLO11* belongs to a family of genes that encode cell surface adhesion molecules. At low oxygen levels, a tandem repeat sequence in this gene changes its copy number, resulting in a serine- and threonine-rich protein, hence causing the cell surface to become more hydrophobic. The more hydrophobic cell surface helps yeast cells attach to each other and form a biofilm at the air-liquid interface, which provides them the oxygen level they need for survival. Another example can be found in RNA enzymes. Hayden and colleagues (2011) showed that RNA enzymes with cryptic variation, that is, populations exposed to several rounds of mutagenesis while keeping their native functions, adapt more rapidly to a new substrate compared to RNA enzyme populations without cryptic variation.

The congruent scenario

The conditions under which genetic perturbations can cause an increase in robustness are very restrictive. They require large populations or high mutation rates (Wagner, 2005). Because environmental perturbations are more frequent and can have high impact on fitness, they can drive the evolution of environmental robustness. Wagner and colleagues (1997) suggested that genetic robustness can emerge as a by-product of selection for environmental robustness. In this case the two forms of robustness are said to be congruent. Examples of correlated genetic and environmental robustness provide evidence for this suggestion. In a striking example, Ancel and Fontana (2000) showed that RNA structures that are robust against thermodynamic perturbations are also robust against mutational perturbations. Further support comes from studies of heat shock proteins, such as Hsp90 and GroEL. These proteins are thought to have evolved to protect organisms from environmental perturbations, but they are found to buffer also against genetic perturbations in Drosophila (Rutherford and Lindquist 1998), plants (Queitsch et al. 2002), and bacteria (Fares et al. 2002).

1.1.3. Mechanistic causes of robustness

Redundancy and degeneracy (also known as distributed robustness) are two basic and prevalent principles that play an important role in achieving robustness in biological systems (de Visser et al. 2003).

Redundancy

Redundancy refers to the coexistence of components with identical functionality such as gene duplicates. Loss of function in one duplicate gene can be compensated by the

other copy (Conant and Wagner 2004; Gu et al. 2003). Moreover, the more similar two duplicates are, the less severe may be the effect of deleting one of them (Conant and Wagner 2004; Gu et al. 2003). A remarkable example of how redundancy causes robustness involves three *thiamin pyrophosphokinase* genes from yeast. All of them encode catalytic subunits of a key protein kinase in cell signaling. Any two of the three genes are dispensable for cell growth (Toda et al. 1987). Kafri and colleagues (2005) provided a mechanistic explanation for such robustness, in which they showed that more essential duplicate genes have functions more similar to each other, and that null-deletion of one copy is often compensated by overexpression of another copy.

Degeneracy

Eliminating duplicates of a gene often causes no severe phenotypic effects, yet thousands of genes whose deletion has no detectable effect are single-copy genes. In line with this, Wagner (2005a) showed that interactions among unrelated genes may be a major cause for mutational robustness. This type of robustness, which emerges through the actions of multiple dissimilar parts, is also called distributed robustness. Specifically, gene knockout experiments and computational work (Edwards and Palsson 2000a; Pál et al. 2006; Segrè et al. 2002) show that in any one environment, many individual reactions of a metabolic network, even reactions in the most central parts of metabolism, such as glycolysis or the citric acid cycle are dispensable. The reason lies in the distributed nature of metabolic systems, where several alternative routes may exist around any blocked pathway. For example, eliminating the first reaction of the pentose phosphate shunt in *E. coli* metabolism has a negligible effect on cell growth, yet it causes large compensatory flux changes elsewhere in metabolism, especially in glycolysis pathway (Edwards and Palsson 1999). In

agreement with this finding, Sauer (2006) showed that these two distinct pathways, glycolysis and the pentose phosphate pathway, can substitute for each other in glucose metabolism. Such alternative metabolic routes can make metabolic networks highly robust to mutations, and ensure that a network continues to produce biosynthetic building blocks and energy carriers.

1.1.4. How to study robustness?

Perturbation Experiments

Empirical evidence for robustness can be gained in multiple ways. The most widely used approach relies on perturbation experiments (Masel and Siegal 2009). In this approach, one perturbs a part of a biological system (a gene), a trait (wing shape) or a capability (glucose synthesis) through mutations. Environmental perturbations in such experiments involve exposure to heat shock (Waddington 1953), mutagenizing agents like ether (Waddington, 1956), or stress conditions such as high salinity (Waddington, 1959). Genetic perturbations include point mutations, loss-of-function or gain-of-function mutations, differential gene regulation through small interfering RNA in key developmental genes, or in other genes of interest (Masel and Siegal 2009). The less a feature's properties change in the face of perturbation, the more robust it is.

Modeling approach

Assessing a system's robustness through perturbation experiments requires many perturbations and subsequent measurements of system properties. This problem is partly alleviated by modeling of biological systems, using both analytical and computational approaches (Blais and Dynlacht 2005; Edwards and Palsson 2000a;

Oberhardt et al. 2009; Price et al. 2002). Such models can provide accurate predictions about a system's robustness, even if systematic perturbations are not feasible.

Gene Regulatory Networks. Models of gene regulatory networks encapsulate interactions between a gene/protein and its regulators (such as proteins, transcription factors, and mRNA). Garg and colleagues (2009) could estimate the robustness of cell differentiation networks to expression noise using a gene regulatory network model. In another example, Ciliberti and colleagues (2007) examined the structure and robustness of millions of transcription regulation networks that regulate both cellular functions and embryonic development in many organisms. They found that radically different network architectures can show the same gene expression pattern.

Metabolic Models. Biochemical network models that represent the metabolism of an organism are also widely used to measure robustness against mutations. A metabolism is a complex chemical reaction system whose metabolic genotype – the DNA encoding the enzymes catalyzing these reactions – can be compactly represented by its complement of metabolic reactions. These constructions are called genome-scale metabolic models and they contain all of the known metabolic reactions in an organism mapped to the genes that encode each enzyme (Becker and Palsson 2005; Feist et al. 2009; Oberhardt et al. 2009).

Metabolisms are highly robust to the elimination of enzyme-coding genes. As exemplified in the Section 1.1.3, loss-of-function mutations in many enzyme-coding genes can leave a metabolic phenotype unaffected (Edwards and Palsson 2000a; Price

et al. 2002). Hence, metabolism can evolve rapidly through mutations that eliminate such genes and through horizontal gene transfer that adds new enzyme-coding genes. This property of metabolism is important for the project I describe in Chapter 2, where, by adding and deleting reactions, I generated random metabolisms, which retain their functions.

To infer metabolic functionality, one can use an important tool for harnessing the knowledge encoded in genome scale metabolic models: Flux Balance Analysis (FBA) (Kauffman et al. 2003; Orth et al. 2010; Smallbone and Simeonidis 2009). FBA uses information about the stoichiometry of reactions in a metabolism to predict the rate at which it can synthesize a given set of molecules (Feist and Palsson 2010). Mutations in metabolic genes affect the activity of enzymes, and thus the rates at which a chemical reaction proceeds. Simulating mutations that will cause metabolic gene deletions and additions in FBA allows the prediction of robustness in metabolic fluxes (that is, the flow rate of metabolites through a network) or in growth rate to perturbations in various enzymes (Edwards and Palsson 2000a; Matias Rodrigues and Wagner 2009). Chapter 2 provides more detail on how FBA works.

1.1.5. Robustness against translation errors

Chapter 3 of this dissertation focuses on robustness to mistranslation and on signatures of selection to minimize the effects of such mutations. Translation is an error-prone process. Mistranslation rates are estimated to occur in every 1,000 to 10,000 translated codons (Ogle and Ramakrishnan 2005a). There are two types of translational errors: (i) missense errors in which an incorrect amino acid is incorporated into a growing peptide chain, and (ii) nonsense errors in which peptide

synthesis terminates prematurely. Both missense and nonsense errors that produce non- and dysfunctional proteins are costly to the cell because they consume amino acids and energy both in their production and during breakdown (Drummond & Wilke, 2009). Accumulation of misfolded or dysfunctional proteins can cause diseases or membrane disruption. Additionally, missense errors may have other effects of large impact. For example, a missense error in a DNA polymerase may temporally increase overall mutation rates (Ninio 1991).

Translationally robust proteins can fold and function properly even if they are mistranslated. Mathematical and computational modeling predicts that this selection pressure will cause proteins to be more thermostable and also to be more tolerant to genetic mutations (Drummond & Wilke, 2008; Drummond, Bloom, Adami, Wilke, & Arnold, 2005; Wilke & Drummond, 2006). One of the first studies to investigate the effect of transcription errors on protein evolution in an experimental system (Goldsmith and Tawfik 2009) confirmed this prediction by showing that a *TEM1 β-lactamase* gene expressed with an error-prone RNA polymerase evolved an increased level of gene expression, increased thermostability and increased mutational robustness.

Translational accuracy

Organisms have evolved various strategies to minimize the effects of mistranslation. Selection for accurate translation is one of those strategies. Akashi (1994) argued that selection for translational accuracy should lead to inhomogeneous codon usage within genes, favoring *optimal codons* that correspond to abundant tRNAs. In line with his argument, Akashi showed that such codons were significantly enriched in more

conserved sites, and also in functionally more important sites (e.g. binding sites) in Drosophila. Further evidence was provided for *E.coli*, where Stoletzki and Eyre-Walker (2007) showed that highly conserved sites and genes have higher codon bias than less conserved ones. Furthermore, they showed that codon bias is positively correlated to gene length and production costs, both indicating selection against missense errors.

Selection for error-mitigation

Genes with optimal codons can still produce large amounts of erroneous peptides. Selection for error-mitigation is suggested to decrease the usage of codons that have a high probability of being mistranslated into radically different amino acids (Archetti, 2004b). Although it does not lead to a reduction of error frequencies, this selection pressure reduces the frequency of the most costly errors at the expense of a larger number of more benign errors, hence providing the translational machinery with robustness against the effect of mistranslation. Pertinent evidence comes from studies on genetic code architecture (Epstein 1966; Woese 1965), which revealed that the structure of the genetic code seems to reduce the effects of mistranslations, because amino acids with similar chemical properties are encoded by similar codons. Further support comes from several studies (Archetti, 2006; Archetti, 2004a, 2004b; Najafabadi, Goodarzi, & Torabi, 2005; Najafabadi, Lehmann, & Omidi, 2007), which show that genes in various organisms tend to "prefer" codons that minimize the effects of mistranslations.

Mistranslation-induced protein misfolding hypothesis

A third strategy to minimize the effects of mistranslation overcomes the disruptive effects of mistranslations on protein structures by increasing the usage of optimal codons at sites where mistranslation is more likely to cause misfolding. Evidence for this strategy is found in multiple organisms, including *E. coli*, yeast, Drosophila and mice (Drummond & Wilke, 2008; Wilke & Drummond, 2006), where translationally optimal codons are more frequently used at sites where mutations are more destabilizing, such as buried amino acids.

1.1.6. Neutral Networks

One common feature of genotype-phenotype interaction is the existence of neutral genotype networks -- connected sets of genotypes that adopt the same phenotype (Ebner, Shackleton, & Shipman, 2001; Wagner, 2008). Any one genotype in such a network can be reached from any other genotype through series of genotypic mutations without altering the phenotype. Examples of neutral networks include RNA sequences that share the same secondary structure (Jörg et al. 2008; Rendel 2011); proteins, where multiple amino acid sequences form the same fold (Bloom et al. 2007; Tóth-Petróczy and Tawfik 2013); regulatory circuits, where many genetically encoded circuit topologies can form the same expression pattern (Ciliberti et al. 2007; Macneil and Walhout 2011), and metabolism, where multiple metabolic genotypes, encoding different combinations of chemical reactions, can confer viability on the same spectrum of nutrients (Von Dassow and Odell 2002; Edwards and Palsson 2000a). Through genotypic changes that do not affect phenotype, vast regions of genotype space can be explored, regions in which molecules with novel phenotypes can lie

(Wagner, 2005; Wagner, 2005b).

1.2. Robustness and Phenotypic Variation

The ability of mutations to bring forth new phenotypes is important for Darwinian evolution. One would think that mutational robustness could only decrease phenotypic variation, because in a robust system, mutations do not easily change a phenotype. However, observations on multiple levels of biological organizations suggest the contrary. Studying the evolution of thermotolerance in an RNA virus, McBride and colleagues (2008) found that populations derived from robust clones evolved greater resistance to heat shock relative to populations founded by non-robust clones. In laboratory evolution experiments, robust proteins evolve new catalytic activities more readily, and proteins with robust folds have evolved a greater diversity of catalytic functions than other proteins (Wagner 2005a). Bloom and colleagues (2006) found that only robust (thermostable) protein variations could tolerate the destabilizing mutations needed to confer novel activities, whereas nonrobust (thermosensitive) proteins could not evolve new activities.

All these studies provide direct empirical evidence that robustness can facilitate phenotypic variation. The reason is that robust phenotypes allow a population to accumulate neutral mutations, increasing genotypic diversity. Because many of these neutral variants harbor distinct phenotypically consequential sensitivities to further genetic modification, mutational robustness can enhance access to phenotypic

variation over time (Wagner, 2008).

1.3. Phenotypic Variation and Its Genetic Determinants

What are the genetic causes of phenotypic variation? There is enormous interest in finding the genetic determinants of phenotypic variation between individuals, populations and species in molecular biology (Frazer et al. 2009; Guryev et al. 2008; Henrichsen et al. 2009b; Sumedha et al. 2007; Tirosh et al. 2006). Broadening our understanding in this regard will not only be valuable for medical genetics but will also provide a better understanding of the phenotypic evolution of complex biological systems. A focus of much research in this area regards the genetic basis of primate evolution (Khaitovich et al. 2006; King and Wilson 1975; Shea 2005; Yang 1998). Striking phenotypic differences exist between humans and their close relatives (Byrne 2000; Shea 2005). Since the release of human (Hattori 2005) and great ape genomes (Locke et al. 2011; Scally et al. 2012; Sequencing and Consortium 2005), numerous genetic differences have been identified at least some of which form the basis for the complex and rapid cultural change that have characterized recent human evolution (Sholtis and Noonan 2010; Varki et al. 2008). Furthermore, it has long been suggested that this phenotypic divergence may be mostly caused by changes in gene regulation (King and Wilson 1975). Because Chapter 4 and 5 of this dissertation focus on expression changes in primates and in human tumors, respectively, I will summarize current knowledge about determinants of such phenotypic variation in the next sections.

1.3.1. Gene regulation

Much of the phenotypic variation is caused by variation in regulatory sequences (Carroll 2000; Romero et al. 2012; Tejedor and Valcárcel 2010; Wray et al. 2003b). In multicellular organisms, for example, gene regulation drives cellular differentiation, leading to the creation of different cell types that possess different gene expression profiles, and hence produce different phenotypes (Barrett, Fletcher, & Wilton, 2012; Carroll, 2000; Choi & Kim, 2008; Tirosh et al., 2006; Wray et al., 2003). Variation in gene expression can therefore help create novel phenotypes.

Regulation of gene expression is a multifaceted process. While the binding of regulatory proteins to the upstream or downstream of a gene can activate, silence or prolong the gene's expression, epigenetic factors such as DNA methylation and histone modifications change the accessibility of the gene to those regulatory proteins (Wray et al., 2003). Genetic regulation can both occur on the transcriptional and translational level. As Chapter 4 and 5 focus on transcriptional regulation, I will briefly introduce the elements of gene expression regulation on the transcriptional level. Following that, I will also discuss translational and epigenetic regulation of gene expression.

Transcriptional regulation

Eukaryotes employ diverse mechanisms to regulate gene expression at the transcriptional level (Barrett et al., 2012; Carroll, 2000; Wray et al., 2003). Changes in transcriptional regulation are an important component of phenotypic variation in physiology, behavior, anatomy, and life history (Burgess, 2013; Spitz & Furlong,

2012; Wray et al., 2003; Wright, Yau, Looseley, & Meyers, 2004). Most genes are differentially transcribed across an organism's life cycle, according to environmental conditions, in different cell types and compartments, and among sexes through diverse regulatory mechanisms (Wray et al., 2003). This is managed through the interplay of many cis-regulatory elements, including promoters, enhancers, and suppressors and proteins that bind to these elements.

At its most fundamental level, the function of a promoter is to integrate information about the status of the cell in which it resides, and to alter the rate of transcriptional initiation of a single gene accordingly. The promoters of genes encoding housekeeping proteins are constitutively active, but they can shut down in response to specific conditions, such as heat shock or starvation (Pirkkala et al. 2001). Other promoters are off by default, but they can be activated in response to specific hormonal, physiological, or environmental cues (Aranda and Pascual 2001).

No consistent sequence motifs exist for promoters of protein-coding genes (Wray et al. 2003a). Two functional features are always present, although they cannot always be recognized from sequence information alone. One is a core promoter, the site upon which the enzymatic machinery of transcription assembles. The other functional feature is a collection of diverse protein binding sites that confer specificity of transcription. Proteins bound to these sites produce a scalar response, the frequency with which new transcripts are initiated (Latchman 1997; Spitz and Furlong 2012; Warren 2002). Proteins in the mitogen-activated protein kinase (MAPK) pathway, for example, control crucial events like cell differentiation, survival and death in eukaryotes, by binding to and regulating the activity of other transcription factors

(Kim et al. 2011; Plotnikov et al. 2011). Dysfunction of this cascade is associated with several types of carcinogenesis (Dhillon et al. 2007). Similarly, the erythroblast transformation-specific (ETS) family, another important family of transcription factors in animal kingdom, is responsible for cell survival and death. Multiple ETS factors become dysfunctional in some disease. For example, the *ETS-related gene* (*ERG*) transcription factor is fused to the *Ewing sarcoma breakpoint region* (*EWS*) in Ewing's sarcoma disease (Sorensen et al. 1994).

Translational regulation

Translational regulation mostly involves controlling the initiation of mRNA translation, which can be modulated by mRNA secondary structure, antisense RNA binding, or protein binding. One of the best known examples of expression regulation through mRNA secondary structure involves the μ gene that encodes immunoglobulin heavy chain. Alteration of its mRNA secondary structure at the ribosome binding site by oligonucleotide replacement mutagenesis revealed a correlation between μ gene's expression levels and accessibility of the ribosome binding site (Wood et al. 1984). In another example, Hüttelmaier and colleagues (2005) carried out an *in vivo* experiment in rabbits to show that the oncofetal protein ZBP1, is involved in the translational repression of β-actin mRNA by blocking translation initiation.

Epigenetic regulation

Epigenetics is the study of heritable genetic changes that are not caused by changes in the DNA sequence (Choi and Kim 2008). Examples of mechanisms that produce such

changes are DNA methylation (Bell et al. 2011; Pai et al. 2011) and histone modification (Thurman et al. 2012; Woo and Li 2012).

DNA methylation is a biochemical process where a methyl group is added to cytosine or adenine DNA nucleotides. The rate of cytosine DNA methylation differs strongly between species: 14% of cytosines are methylated in plants, 8% in mice, 2.3% in *E. coli*, 0.03% in fruit fly, and virtually none in yeast (Capuano et al. 2014). In mammals, genes can be methylated from CpG islands in the upstream (Deaton and Bird 2011), where tandem cytosine and guanine nucleotides are clustered together, which can help repress gene expression. High methylation of gene promoters correlates with low or no transcription (Bell et al. 2011; Suzuki and Bird 2008).

DNA methylation is essential for normal development (Bergman and Cedar 2013). For example, Oct-4, an important transcription factor responsible for self-renewal of undifferentiated embryonic stem cells is silenced by hypermethylation during differentiation (Feldman et al. 2006). Abnormal methylation is associated with a number of diseases, such as cancer and atherosclerosis (Bergman and Cedar 2013; Robertson 2005). In cancer, CpG sites in gene promoters acquire abnormal hypermethylation, which results in silencing of tumor-suppressor genes, such as *Adenomatous polyposis coli* (*APC*) and *Breast cancer 1, early onset* (*BRCA1*), and of genes responsible for DNA repair, such as DNA mismatch repair gene, *MLH1* (Taberlay and Jones 2011).

The histone proteins of eukaryotic cells package and order the DNA into structural units called nucleosomes. A huge catalogue of histone modifications have been

described, such as acetylation, methylation and phosphorylation of various amino acids on histones, most of which lead to transcriptional activation of the nearby genes (Gaffney et al. 2012; Woo and Li 2012). For example, switch/sucrose nonfermentable (SWI/SNF) complex in yeast destabilizes DNA-histone interactions and opens up chromatin structure, thereby allowing the transcription of genes located within the chromatin (Whitehouse et al. 1999). In another example, Polycomb proteins silence the expression of *Hox* genes, which encode a group of proteins responsible for body development by changing their chromatin structure in Drosophila (Stankunas et al. 1998).

1.3.2. Genetic mutations responsible for phenotypic variation

Identifying DNA mutations responsible for phenotypic variation is one goal of evolutionary genetics. Genetic mutations either alter single nucleotides or create insertions and deletions. These inserted or deleted sequences can be very short, and comprise only couple of nucleotides, such as in variable tandem repeats (Gemayel et al. 2010), or very long such as in chromosomal copy number variation (Redon et al. 2006). Genes themselves can also be deleted or duplicated (Zhang 2003). Rates for these different kinds of mutations vary greatly among organism and among types of mutations (Kumar and Subramanian 2002; Lynch 2010; Scally and Durbin 2012). For example, in humans point mutations are estimated to occur at a rate of 10^{-8} per nucleotide per generation (Scally and Durbin 2012), whereas mutation rates of tandem repeat sequences lie between 10^{-3} and 10^{-7} per cell division (Fan and Chu 2007; Legendre et al. 2007).

Single nucleotide polymorphisms

The four-winged fly that results from short nucleotide mutations in *Ubx* gene promoter in Drosophila is perhaps the most remarkable example (Simon et al. 1990) of how small genotypic alterations can change the phenotype. An important class of them, single-nucleotide polymorphisms (SNPs) have long been known to be associated with phenotypic variation (Kruglyak 1999; Shastry 2002; Stranger et al. 2007a; Wang and Moult 2001). Another famous but more recent example comes from the Tibetan population, where multiple SNP's in the genes *Endothelial PAS domain 1* and *Hypoxia-inducible factor prolyl hydroxylase 1*, which are involved in the low oxygen response confer high altitude adaptation to this population (Peng et al. 2011; Yi et al. 2010). Another study on multiple human populations (Li et al. 2010) showed that SNPs located in untranslated regions of 18,000 genes elevate expression variation between populations.

Large copy number variants

During the last few years, copy number variants of DNA segments that are one kilobase or larger in size have attracted much attention. They are quite prevalent in eukaryotes: roughly 10 per cent of many eukaryotic genomes consist of such segmental duplications (Henrichsen et al. 2009a; Jakobsson et al. 2008; Mills et al. 2011). Copy number variants can have dramatic phenotypic consequences (Chaignat et al. 2011; Henrichsen et al. 2009b; Stranger et al. 2007a; Wang et al. 2011; Zhou et al. 2011). One example can be found in zebrafish, where copy number amplification of regions containing the *mannose-binding lectin* gene influence its susceptibility to bacterial infection (Jackson et al. 2007). Another example involves large structural

rearrangements in multiple chromosomes that are found to be strongly associated with schizophrenia in human (Stefansson et al. 2008). Multiple studies show that increased copy number can be positively (McCarroll 2008; Somerville et al. 2005) or negatively (Lee et al. 2006a) correlated with gene expression levels.

Gene duplication

Gene duplication is an important kind of genotypic variation for creating new phenotypes in organisms. Lynch and Conery (2000) estimated that gene duplications arise and get fixed at an approximate rate of 10^{-8} per gene per genome in eukaryotes. Many novel gene functions have evolved through gene duplication, which has contributed tremendously to the evolution of developmental programs in various organisms. Gene duplication is associated with increased gene expression divergence and morphological diversification (Conant and Wolfe 2008; Dong et al. 2011; Hanada et al. 2009; Magadum et al. 2013). The evolution of the antifreeze protein in *Antarctic zoarcid* fish provides a prime example of phenotypic variation conferred by gene duplication (Deng et al. 2010). After a duplication event of the *sialic acid synthase* gene in *Antarctic zoarcid* fish, one copy accumulated several mutations to gain antifreeze functionality, allowing the fish to survive in the frigid temperatures of the Antarctic Seas. A remarkable example of how gene duplication causes expression variation comes from a comparison of recent human and chimpanzee duplications (Cheng et al., 2005), which revealed that more than half of the human-specific duplicates show a significant overexpression in the human lineage.

Tandem repeat instability

Another class of copy number variation involves short sequences up to 50 or 100 nucleotides that are repeated tandemly. They are of a special importance for this dissertation as two of the following chapters (Chapter 4 and 5) are devoted to their phenotypic consequences for gene expression levels.

Tandem repeats are extremely unstable. Their copy number varies 10 to 100,000 times more frequent than other parts of genome in eukaryotes (Gemayel et al. 2010). The reason is that tandem repeats are prone to an error called strand slippage, which occurs predominantly during cell replication when there is a mispairing between the template and complementary DNA strands (Levinson and Gutman 1987). When the newly synthesized strand denatures from the template strand during synthesis of the tandem repeat sequence, it can occasionally pair with another part of the repeat sequence due to self-compatibility. If the template strand is looped out, then tandem repeat copy number decreases. If the complementary strand loops out, copy number increases.

Copy number changes in tandem repeats often have tremendous phenotypic consequences. For example, unstable repeats located in or near human genes can lead to neurodegenerative diseases such as Huntington disease and muscular dystrophy (Bates 2005). Apart from their role in disease, variable repeats can confer non-pathogenic phenotypic variation. A compelling example comes from a study on dog skull morphology (Fondon and Garner 2004), which compared genomic and morphological data from different dog breeds and revealed immense morphological changes caused by variable tandem repeats in two developmental genes, *Alx-4* and

Runx-2. Another striking example can be found in *FLO1*, a cell surface adhesion gene in yeast. Experimentally altering the copy number of a 100 nucleotide long repeat region, which is polymorphic between various yeast strains, changes cell adherence, thereby facilitating adaptation to different environments (Verstrepen et al. 2005).

Vinces and colleagues (2009) showed that changing the copy number of tandem repeats in a gene's promoter has a direct effect on its expression level in yeast species. They also demonstrated that genes regulated by repeat-containing promoters show significantly higher rates of transcriptional variation. A striking example can be found in tilapia, an important aquacultural fish. Variable CA repeats in the promoter of the *prl1* gene, which encodes a hormone involved in osmoregulation, show an association with both *prl1* expression as well as the fish's response to salt stress (Streelman and Kocher 2002). Variation in a tandem repeat can modulate gene expression also by changing the copy number of binding sites for regulatory proteins, such as transcription factors. For example, the tumor suppressor p53 activates the transcription of *PIG3*, a gene involved in p53-mediated cell death, by interacting with a pentanucleotide tandem repeat sequence in the promoter of this gene (Contente et al. 2002). Different copy numbers of the repeat sequence associate with different expression levels of the *PIG3* gene.

Tandem repeat instability and cancer

Tumorigenesis is partly driven by mutations that increase genomic repeat instability by allowing tumor cells to rapidly acquire various mutations required for cellular transformation through an increase in random mutation events (Aguilera and García-Muse 2013; Maslov and Vijg 2009). Genomic instability can originate from

deficiencies in the DNA mismatch repair system (Hewish et al. 2010; Woerner et al. 2003; Zienolddiny et al. 1999), which allow DNA damage to accumulate, and give rise to further mutations, especially in tandem repeats (Cancer and Atlas 2012; Fearon 2011; Gurin et al. 1999). For example, multiple studies (Bubb et al. 1996; Cancer and Atlas 2012; Fearon 2011; Hewish et al. 2010) have shown that inactivation of one of several mismatch repair genes is responsible for increased tandem repeat instability seen in more than 90 per cent of hereditary non-polyposis colorectal cancers, and in 15 per cent of non-hereditary colorectal cancers. This high genotypic variation increases the probability of tumors harboring a therapy-resistant phenotype and has been hypothesized to endow tumors with the necessary adaptability to survive and recur after treatment (Kitano 2004; Tischfield and Shao 2003). In Chapter 5, I describe how tandem repeat instability can cause tumor-specific gene expression changes.

1.4. Thesis Outline

This dissertation covers the work of four years, in which I studied phenotypic variation and robustness on various levels of biological organizations. Each chapter is devoted to one of these projects. The first research chapter (**Chapter 2)** describes a computational analysis that focuses on the tradeoffs between different metabolic network properties that contribute to phenotypic variation in a synthetic microbial metabolism. These properties include the number of carbon sources a metabolism can use to survive, the number of different molecules a metabolism can synthesize, the

number of actively used reactions in a metabolism, and how much waste a metabolism produces. Variations in these properties explain most of the variation in the biomass synthesis rates of a metabolism. Furthermore, I show that biochemically related molecules (e.g. amino acids) can be synthesized at higher rates, because their synthesis produces less waste. The observations in this study are relevant for synthetic metabolism design, which has become possible thanks to ongoing advances in sequencing and de-novo synthesis of DNA (Cheng and Lu 2012; Nandagopal and Elowitz 2011; Smolke and Silver 2011).

In **Chapter 3**, I ask to what extent codon changes caused by mutation or mistranslation may affect physicochemical amino acid properties or protein folding. I find that codons of ligand-binding amino acids are on average more robust to errors than those of non-binding amino acids. Selection for error mitigation at the translational level can be responsible for this phenomenon. The finding of this study suggests that natural selection can affect the robustness of very small units of biological organization.

In **Chapter 4**, I focus on how a particular genotypic instability, tandem repeat variation relates to gene expression divergence in primates. By doing so, I find that genes with tandem repeats in their regulatory regions have significantly higher expression divergence. Similarly, I show that human gene duplicates with tandem repeats diverge in expression more than duplicates without tandem repeats. Hence, tandem repeats, far from just being a source of genetic diseases, may contribute substantially to the divergence of gene expression in primates.

Since tandem repeat instability is a hallmark of colorectal tumors, I study in **Chapter 5** the phenotypic consequences of tandem repeat instability between tumor and normal tissues of the same individual. I first show that tumor genomes are enriched for repeat instability, i.e. de novo repeats, repeat loss, and copy number variation. I then show that genes with repeat instability are significantly overexpressed, also in well-studied cancer pathways. These findings suggest an important role for promoter tandem repeat instability in differential gene expression of colorectal tumors.

1.5. References

Aguilera, A., & García-Muse, T. (2013). Causes of genome instability. *Annual Review of Genetics*, *47*, 1–32. doi:10.1146/annurev-genet-111212-133232

Akashi, H. (1994). Synonymous Codon Usage in Drosophila Melanogaster: Natural Selection and Translational Accuracy. *Genetics*, *136*(3), 927–935. Retrieved from http://www.genetics.org/cgi/content/abstract/136/3/927

Ancel, L. W., & Fontana, W. (2000). Plasticity, evolvability, and modularity in RNA. *Journal of Experimental Zoology*, *288*, 242–283. doi:10.1002/1097-010X(20001015)288:3<242::AID-JEZ5>3.0.CO;2-O

Aranda, A., & Pascual, A. (2001). Nuclear hormone receptors and gene expression. *Physiological Reviews*, *81*, 1269–1304.

Archetti, M. (2004a). Codon usage bias and mutation constraints reduce the level of error minimization of the genetic code. *Journal of Molecular Evolution*, *59*(2), 258–66. doi:10.1007/s00239-004-2620-0

Archetti, M. (2004b). Selection on codon usage for error minimization at the protein level. *Journal of Molecular Evolution*, *59*(3), 400–15. doi:10.1007/s00239-004-2634-7

Archetti, M. (2006). Genetic robustness and selection at the protein level for synonymous codons. *Journal of Evolutionary Biology*, *19*(2), 353–65. doi:10.1111/j.1420-9101.2005.01029.x

Barrett, L. W., Fletcher, S., & Wilton, S. D. (2012). Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. *Cellular and Molecular Life Sciences CMLS*, 1–22. doi:10.1007/s00018-012-0990-9

Batada, N. N., & Hurst, L. D. (2007). Evolution of chromosome organization driven by selection for reduced gene expression noise. *Nature Genetics*, *39*, 945–949. doi:10.1038/ng2071

Bates, G. P. (2005). The molecular genetics of Huntington disease — a history. *Nature*, *6*, 766–773.

Becker, S. a, & Palsson, B. Ø. (2005). Genome-scale reconstruction of the metabolic network in Staphylococcus aureus N315: an initial draft to the two-dimensional annotation. *BMC Microbiology*, *5*, 8. doi:10.1186/1471-2180-5-8

Bell, J. T., Pai, A. a, Pickrell, J. K., Gaffney, D. J., Pique-Regi, R., Degner, J. F., … Pritchard, J. K. (2011). DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biology*, *12*(1), R10. doi:10.1186/gb-2011-12-1-r10

Benfey, P. N., & Mitchell-Olds, T. (2008). From genotype to phenotype: systems biology meets natural variation. *Science (New York, N.Y.)*, *320*, 495–497. doi:10.1126/science.1153716

Bergman, Y., & Cedar, H. (2013). DNA methylation dynamics in health and disease. *Nature Structural & Molecular Biology*, *20*, 274–281. doi:10.1038/nsmb.2518

Blais, A., & Dynlacht, B. D. (2005). Constructing transcriptional regulatory networks. *Genes & Development*, *19*, 1499–1511. doi:10.1101/gad.1325605

Bloom, J. D., Labthavikul, S. T., Otey, C. R., & Arnold, F. H. (2006). Protein stability promotes evolvability. *Proceedings of the National Academy of Sciences of the United States of America*, *103*, 5869–5874. doi:10.1073/pnas.0510098103

Bloom, J. D., Romero, P. A., Lu, Z., & Arnold, F. H. (2007). Neutral genetic drift can alter promiscuous protein functions, potentially aiding functional evolution. *Biology Direct*, *2*, 17. doi:10.1186/1745-6150-2-17

Bubb, V. J., Curtis, L. J., Cunningham, C., Dunlop, M. G., Carothers, A. D., Morris, R. G., … Wyllie, A. H. (1996). Microsatellite instability and the role of hMSH2 in sporadic colorectalcancer. *Oncogene*, *12*(12), 2641–2649. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=8700523

Bull, J. (1987). Evolution of phenotypic variance. *Evolution*, *41*, 303–315. doi:10.2307/2409140

Burgess, D. J. (2013). Gene expression: colorectal cancer classifications. *Nature Reviews. Cancer*, *13*, 380–1. doi:10.1038/nrc3529

Byrne, R. (2000). Evolution of primate cognition. *Cognitive Science*, *24*, 543–570. doi:10.1207/s15516709cog2403_8

Cancer, T., & Atlas, G. (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, *487*, 330–7. doi:10.1038/nature11252

Capuano, F., Mülleder, M., Kok, R., Blom, H., & Ralser, M. (2014). Cytosine DNA Methylation Is Found in Drosophila melanogaster but Absent in Saccharomyces cerevisiae, Schizosaccharomyces pombe, and Other Yeast Species. *Anal. Chem.*, *86*, 3697–702. doi:10.1021/ac500447w

Carroll, S. B. (2000). Endless forms: the evolution of gene regulation and morphological diversity. *Cell*, *101*, 577–580. doi:10.1016/S0092-8674(00)80868-5

Chaignat, E., Yahya-Graison, E. A., Henrichsen, C. N., Chrast, J., Schütz, F., Pradervand, S., & Reymond, A. (2011). Copy number variation modifies expression time courses. *Genome Research*, *21*, 106–113. doi:10.1101/gr.112748.110

Cheng, A. A., & Lu, T. K. (2012). Synthetic Biology: An Emerging Engineering Discipline. *Annual Review of Biomedical Engineering*. doi:10.1146/annurev-bioeng-071811-150118

Cheng, Z., Ventura, M., She, X., Khaitovich, P., Graves, T., Osoegawa, K., … Eichler, E. E. (2005). A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature*, *437*, 88–93. doi:10.1038/nature04000

Choi, J. K., & Kim, Y.-J. (2008). Epigenetic regulation and the variability of gene expression. *Nature Genetics*, *40*(2), 141–147. doi:10.1038/ng.2007.58

Ciliberti, S., Martin, O. C., & Wagner, A. (2007). Robustness Can Evolve Gradually in Complex Regulatory Gene Networks with Varying Topology. *PLoS Computational Biology*, *3*(2), 10. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17274682

Conant, G. C., & Wagner, A. (2004). Duplicate genes and robustness to transient gene knock-downs in Caenorhabditis elegans. *Proceedings. Biological Sciences / The Royal Society*, *271*, 89–96. doi:10.1098/rspb.2003.2560

Conant, G. C., & Wolfe, K. H. (2008). Turning a hobby into a job: How duplicated genes find new functions. *Nature Reviews Genetics*, *9*(12), 938–950. doi:10.1038/nrg2482

Contente, A., Dittmer, A., Koch, M. C., Roth, J., & Dobbelstein, M. (2002). A polymorphic microsatellite that mediates induction of PIG3 by p53. *Nature Genetics*, *30*, 315–320. doi:10.1038/ng836

De Visser, J. A. G. M., Hermisson, J., Wagner, G. P., Ancel Meyers, L., Bagheri-Chaichian, H., Blanchard, J. L., … Whitlock, M. C. (2003). Perspective: Evolution and detection of genetic robustness. *Evolution; International Journal of Organic Evolution*, *57*, 1959–1972. doi:10.1554/02-750R

Deaton, A. M., & Bird, A. (2011). CpG islands and the regulation of transcription. *Genes & Development*, *25*, 1010–1022. doi:10.1101/gad.2037511

Deng, C., Cheng, C.-H. C., Ye, H., He, X., & Chen, L. (2010). Evolution of an antifreeze protein by neofunctionalization under escape from adaptive conflict. *Proceedings of the National Academy of Sciences of the United States of America*, *107*, 21593–21598. doi:10.1073/pnas.1007883107

Dhillon, A. S., Hagan, S., Rath, O., & Kolch, W. (2007). MAP kinase signalling pathways in cancer. *Oncogene*, *26*, 3279–3290. doi:10.1038/sj.onc.1210421

Dong, D., Yuan, Z., & Zhang, Z. (2011). Evidences for increased expression variation of duplicate genes in budding yeast: from cis- to trans-regulation effects. *Nucleic Acids Research*, *39*(3), 837–847. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3035465&tool=pmcentrez&rendertype=abstract

Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O., & Arnold, F. H. (2005). Why highly expressed proteins evolve slowly. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(40), 14338–14343. Retrieved from http://arxiv.org/abs/q-bio/0506002

Drummond, D. A., & Wilke, C. O. (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*, *134*(2), 341–352. doi:10.1016/j.cell.2008.05.042

Drummond, D. A., & Wilke, C. O. (2009). The evolutionary consequences of erroneous protein synthesis. *Nature Reviews Genetics*, *10*(10), 715–724. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19763154

Ebner, M., Shackleton, M., & Shipman, R. (2001). How neutral networks influence evolvability. *Complexity*, *7*, 19–33. doi:10.1002/cplx.10021

Edwards, J. S., & Palsson, B. O. (2000). Robustness analysis of the Escherichia coli metabolic network. *Biotechnology Progress*, *16*(6), 927–39. doi:10.1021/bp0000712

Edwards, J. S., & Palsson, B. Ø. (1999). Systems properties of the Haemophilus influenzae Rd metabolic genotype. *Journal of Biological Chemistry*, *274*(25), 17410–17416. doi:10.1074/jbc.274.25.17410

Epstein, C. J. (1966). Role of the amino-acid "code" and of selection for conformation in the evolution of proteins. *Nature*, *210*(5031), 25–28. Retrieved from http://adsabs.harvard.edu/abs/1966Natur.210...25E

Fan, H., & Chu, J.-Y. (2007). A Brief Review of Short Tandem Repeat Mutation. *Genomics, Proteomics & Bioinformatics*. doi:10.1016/S1672-0229(07)60009-6

Fares, M. A., Ruiz-González, M. X., Moya, A., Elena, S. F., & Barrio, E. (2002). Endosymbiotic bacteria: groEL buffers against deleterious mutations. *Nature*, *417*, 398. doi:10.1038/417398a

Fearon, E. R. (2011). Molecular genetics of colorectal cancer. *Annual Review of Pathology*, *6*, 479–507. doi:10.1146/annurev-pathol-011110-130235

Feist, A. M., Herrgård, M. J., Thiele, I., Reed, J. L., & Palsson, B. Ø. (2009). Reconstruction of biochemical networks in microorganisms. *Nature Reviews. Microbiology*, *7*(2), 129–43. doi:10.1038/nrmicro1949

Feist, A. M., & Palsson, B. O. (2010). The biomass objective function. *Current Opinion in Microbiology*, *13*(3), 344–349. doi:10.1016/j.mib.2010.03.003

Feldman, N., Gerson, A., Fang, J., Li, E., Zhang, Y., Shinkai, Y., … Bergman, Y. (2006). G9a-mediated irreversible epigenetic inactivation of Oct-3/4 during early embryogenesis. *Nature Cell Biology*, *8*, 188–194. doi:10.1038/ncb1353

Fidalgo, M., Barrales, R. R., Ibeas, J. I., & Jimenez, J. (2006). Adaptive evolution by mutations in the FLO11 gene. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(30), 11228–11233. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1544070&tool=pmcentrez&rendertype=abstract

Fisher, R. A. (1928). The possible modifications of the response of the wild type to recurrent mutations. *The American Naturalist*, *62*(679), 115–126.

Fondon, J. W., & Garner, H. R. (2004). Molecular origins of rapid and continuous morphological evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(52), 18058–18063. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=539791&tool=pmcentrez&rendertype=abstract

Frazer, K. A., Murray, S. S., Schork, N. J., & Topol, E. J. (2009). Human genetic variation and its contribution to complex traits. *Nature Reviews. Genetics*, *10*, 241–251. doi:10.1038/nrg2554

Gaffney, D. J., McVicker, G., Pai, A. a., Fondufe-Mittendorf, Y. N., Lewellen, N., Michelini, K., … Pritchard, J. K. (2012). Controls of Nucleosome Positioning in the Human Genome. *PLoS Genetics*, *8*(11), e1003036. doi:10.1371/journal.pgen.1003036

Garg, A., Mohanram, K., Di Cara, A., De Micheli, G., & Xenarios, I. (2009). Modeling stochasticity and robustness in gene regulatory networks. In *Bioinformatics* (Vol. 25). doi:10.1093/bioinformatics/btp214

Gemayel, R., Vinces, M. D., Legendre, M., & Verstrepen, K. J. (2010). Variable Tandem Repeats Accelerate Evolution of Coding and Regulatory Sequences. *Annual Review of Genetics*. doi:10.1146/annurev-genet-072610-155046

Goldsmith, M., & Tawfik, D. S. (2009). Potential role of phenotypic mutations in the evolution of protein expression and stability. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 6197–6202. doi:10.1073/pnas.0809506106

Gu, Z., Steinmetz, L., Gu, X., Scharfe, C., Davis, R., & Li, W. (2003). Role of duplicate genes in genetic robustness against null mutations. *Nature*, 63–66. doi:10.1038/nature01226.1.

Guo, H. H., Choe, J., & Loeb, L. A. (2004). Protein tolerance to random amino acid change. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(25), 9205–9210. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=438954&tool=pmcentrez&rendertype=abstract

Gurin, C. C., Federici, M. G., Kang, L., & Boyd, J. (1999). Causes and consequences of microsatellite instability in endometrial carcinoma. *Cancer Research*, *59*(2), 462–466. Retrieved from http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=m&form=6&dopt=r&uid=0009927063

Guryev, V., Saar, K., Adamovic, T., Verheul, M., Van Heesch, S. A., Cook, S., … Cuppen, E. (2008). Distribution and functional impact of DNA copy number variation in the rat. *Nat Genet*, *40*(5), 538–545.

Hanada, K., Kuromori, T., Myouga, F., Toyoda, T., & Shinozaki, K. (2009). Increased Expression and Protein Divergence in Duplicate Genes Is Associated with Morphological Diversification. *PLoS Genetics*, *5*(12), 7. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20041196

Hattori, M. (2005). Finishing the euchromatic sequence of the human genome. *Tanpakushitsu Kakusan Koso Protein Nucleic Acid Enzyme*, *50*(2), 162–168. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15704464

Hayden, E. J., Ferrada, E., & Wagner, A. (2011). Cryptic genetic variation promotes rapid evolutionary adaptation in an RNA enzyme. *Nature*, *474*, 92–95. doi:10.1038/nature10083

Henrichsen, C. N., Chaignat, E., & Reymond, A. (2009). Copy number variants, diseases and gene expression. *Human Molecular Genetics*. doi:10.1093/hmg/ddp011

Henrichsen, C. N., Vinckenbosch, N., Zöllner, S., Chaignat, E., Pradervand, S., Schütz, F., … Reymond, A. (2009). Segmental copy number variation shapes tissue transcriptomes. *Nature Genetics*, *41*(4), 424–9. doi:10.1038/ng.345

Hewish, M., Lord, C. J., Martin, S. A., Cunningham, D., & Ashworth, A. (2010). Mismatch repair deficient colorectal cancer in the era of personalized treatment. *Nature Reviews. Clinical Oncology*, *7*, 197–208. doi:10.1038/nrclinonc.2010.18

Huang, W., Petrosino, J., Hirsch, M., Shenkin, P. S., & Palzkill, T. (1996). Amino acid sequence determinants of beta-lactamase structure and activity. *Journal of Molecular Biology*, *258*(4), 688–703. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8637002

Hüttelmaier, S., Zenklusen, D., Lederer, M., Dictenberg, J., Lorenz, M., Meng, X., … Singer, R. H. (2005). Spatial regulation of beta-actin translation by Src-dependent phosphorylation of ZBP1. *Nature*, *438*, 512–515. doi:10.1038/nature04115

Jackson, A. N., McLure, C. A., Dawkins, R. L., & Keating, P. J. (2007). Mannose binding lectin (MBL) copy number polymorphism in Zebrafish (D. rerio) and identification of haplotypes resistant to L. anguillarum. *Immunogenetics*, *59*, 861–872. doi:10.1007/s00251-007-0251-5

Jakobsson, M., Scholz, S. W., Scheet, P., Gibbs, J. R., VanLiere, J. M., Fung, H.-C., … Singleton, A. B. (2008). Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*, *451*(7181), 998–1003. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18288195

Jörg, T., Martin, O. C., & Wagner, A. (2008). Neutral network sizes of biological RNA molecules can be computed and are not atypically small. *BMC Bioinformatics*, *9*, 464. doi:10.1186/1471-2105-9-464

Kafri, R., Bar-Even, A., & Pilpel, Y. (2005). Transcription control reprogramming in genetic backup circuits. *Nature Genetics*, *37*, 295–299. doi:10.1038/ng1523

Kauffman, K. J., Prakash, P., & Edwards, J. S. (2003). Advances in flux balance analysis. *Current Opinion in Biotechnology*, *14*(5), 491–496. doi:10.1016/j.copbio.2003.08.001

Khaitovich, P., Enard, W., Lachmann, M., & Pääbo, S. (2006). Evolution of primate gene expression. *Nature Reviews. Genetics*, *7*(9), 693–702. doi:10.1038/nrg1940

Kim, Y., Andreu, M. J., Lim, B., Chung, K., Terayama, M., Jim??nez, G., … Shvartsman, S. Y. (2011). Gene regulation by MAPK substrate competition. *Developmental Cell*, *20*, 880–887. doi:10.1016/j.devcel.2011.05.009

King, M. C., & Wilson, A. C. (1975). Evolution at two levels in humans and chimpanzees. *Science*, *188*(4184), 107–116. doi:10.1126/science.1090005

Kitano, H. (2004). Biological robustness. *Nature Reviews. Genetics*, *5*, 826–837. doi:10.1109/SICE.2008.4654600

Kruglyak, L. (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics*, *22*, 139–144. doi:10.1038/9642

Kumar, S., & Subramanian, S. (2002). Mutation rates in mammalian genomes. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(2), 803–8. doi:10.1073/pnas.022629899

Ladbury, J. E., & Arold, S. T. (2012). Noise in cellular signaling pathways: Causes and effects. *Trends in Biochemical Sciences*. doi:10.1016/j.tibs.2012.01.001

Latchman, D. S. (1997). Transcription factors: An overview. *International Journal of Biochemistry and Cell Biology*. doi:10.1016/S1357-2725(97)00085-X

Lee, J. A., Madrid, R. E., Sperle, K., Ritterson, C. M., Hobson, G. M., Garbern, J., … Inoue, K. (2006). Spastic paraplegia type 2 associated with axonal neuropathy and apparent PLP1 position effect. *Annals of Neurology*, *59*, 398–403. doi:10.1002/ana.20732

Legendre, M., Pochet, N., Pak, T., & Verstrepen, K. J. (2007). Sequence-based estimation of minisatellite and microsatellite repeat variability. *Genome Research*, *17*(12), 1787–1796. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2099588&tool=pmcentrez&rendertype=abstract

Lehner, B. (2013). Genotype to phenotype: lessons from model organisms for human genetics. *Nat Rev Genet*, *14*, 168–178. doi:10.1038/nrg3404

Levinson, G., & Gutman, G. A. (1987). Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Molecular Biology and Evolution*, *4*(3), 203–221. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=3328815

Li, J., Liu, Y., Kim, T., Min, R., & Zhang, Z. (2010). Gene Expression Variability within and between Human Populations and Implications toward Disease Susceptibility. *PLoS Computational Biology*, *6*(8), 10. Retrieved from http://dx.plos.org/10.1371/journal.pcbi.1000910

Li, X., Cassidy, J. J., Reinke, C. A., Fischboeck, S., & Carthew, R. W. (2009). A MicroRNA Imparts Robustness against Environmental Fluctuation during Development. *Cell*, *137*, 273–282. doi:10.1016/j.cell.2009.01.058

Locke, D. P., Hillier, L. W., Warren, W. C., Worley, K. C., Nazareth, L. V, Muzny, D. M., … Wilson, R. K. (2011). Comparative and demographic analysis of orang-utan genomes. *Nature*, *469*, 529–533. doi:10.1038/nature09687

Lynch, M. (2010). Evolution of the mutation rate. *Trends in Genetics TIG*, *26*(8), 345–352. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2910838&tool=pmcentrez&rendertype=abstract

Lynch, M., & Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science (New York, N.Y.)*, *290*, 1151–1155. doi:10.1126/science.290.5494.1151

Macneil, L. T., & Walhout, A. J. M. (2011). Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. *Genome Research*, *21*(5), 645–57. doi:10.1101/gr.097378.109

Magadum, S., Banerjee, U., Murugan, P., Gangapur, D., & Ravikesavan, R. (2013). Gene duplication as a major force in evolution. *Journal of Genetics*, *92*, 155–161. doi:10.1007/s12041-013-0212-8

Masel, J., & Siegal, M. L. (2009). Robustness: mechanisms and consequences. *Trends in Genetics*. doi:10.1016/j.tig.2009.07.005

Maslov, A. Y., & Vijg, J. (2009). Genome instability, cancer and aging. *Biochimica et Biophysica Acta - General Subjects*. doi:10.1016/j.bbagen.2009.03.020

Materon, I. C., & Palzkill, T. (2001). Identification of residues critical for metallo-beta-lactamase function by codon randomization and selection. *Protein Science : A Publication of the Protein Society*, *10*, 2556–2565. doi:10.1110/ps.40884

Matias Rodrigues, J. F., & Wagner, A. (2009). Evolutionary plasticity and innovations in complex metabolic reaction networks. *PLoS Computational Biology*, *5*(12), e1000613. doi:10.1371/journal.pcbi.1000613

McBride, R. C., Ogbunugafor, C. B., & Turner, P. E. (2008). Robustness promotes evolvability of thermotolerance in an RNA virus. *BMC Evolutionary Biology*, *8*, 231. doi:10.1186/1471-2148-8-231

McCarroll, S. A. (2008). Extending genome-wide association studies to copy-number variation. *Human Molecular Genetics*, *17*(R2), R135–R142. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18852202

Mills, R. E., Walter, K., Stewart, C., Handsaker, R. E., Chen, K., Alkan, C., … Korbel, J. O. (2011). Mapping copy number variation by population-scale genome sequencing. *Nature*, *470*, 59–65. doi:10.1038/nature09708

Mitchell-Olds, T., Willis, J. H., & Goldstein, D. B. (2007). Which evolutionary processes influence natural genetic variation for phenotypic traits? *Nature Reviews. Genetics*, *8*, 845–856. doi:10.1038/nrg2207

Najafabadi, H. S., Goodarzi, H., & Torabi, N. (2005). Optimality of codon usage in Escherichia coli due to load minimization. *Journal of Theoretical Biology*, *237*(2), 203–209. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15932760

Najafabadi, H. S., Lehmann, J., & Omidi, M. (2007). Error minimization explains the codon usage of highly expressed genes in Escherichia coli. *Gene*, *387*(1-2), 150–155. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17097242

Nandagopal, N., & Elowitz, M. B. (2011). Synthetic biology: integrated gene circuits. *Science (New York, N.Y.)*, *333*, 1244–1248. doi:10.1126/science.1207084

Newman, J. R. S., Ghaemmaghami, S., Ihmels, J., Breslow, D. K., Noble, M., DeRisi, J. L., & Weissman, J. S. (2006). Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature*, *441*, 840–846. doi:10.1038/nature04785

Ninio, J. (1991). Transient mutators: a semiquantitative analysis of the influence of translation and transcription errors on mutation rates. *Genetics*, *129*, 957–962.

Oberhardt, M. A., Palsson, B. Ø., & Papin, J. A. (2009). Applications of genome-scale metabolic reconstructions. *Molecular Systems Biology*, *5*(320), 320. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19888215

Ogle, J. M., & Ramakrishnan, V. (2005). Structural insights into translational fidelity. *Annual Review of Biochemistry*, *74*, 129–177. doi:10.1146/annurev.biochem.74.061903.155440

Oliveira, M. M., Shingleton, A. W., & Mirth, C. K. (2014). Coordination of Wing and Whole-Body Development at Developmental Milestones Ensures Robustness against Environmental and Physiological Perturbations. *PLoS Genetics*, *10*(6).

Orth, J. D., Thiele, I., & Palsson, B. Ø. (2010). What is flux balance analysis? *Nature Biotechnology*, *28*(3), 245–8. doi:10.1038/nbt.1614

Pai, A. a., Bell, J. T., Marioni, J. C., Pritchard, J. K., & Gilad, Y. (2011). A Genome-Wide Study of DNA Methylation Patterns and Gene Expression Levels in Multiple Human and Chimpanzee Tissues. *PLoS Genetics*, *7*(2), e1001316. doi:10.1371/journal.pgen.1001316

Pál, C., Papp, B., Lercher, M. J., Csermely, P., Oliver, S. G., & Hurst, L. D. (2006). Chance and necessity in the evolution of minimal metabolic networks. *Nature*, *440*(7084), 667–70. doi:10.1038/nature04568

Pearson, J. C., Lemons, D., & McGinnis, W. (2005). Modulating Hox gene functions during animal body patterning. *Nature Reviews. Genetics*, *6*, 893–904. doi:10.1038/nrg1726

Peng, Y., Yang, Z., Zhang, H., Cui, C., Qi, X., Luo, X., … Su, B. (2011). Genetic variations in tibetan populations and high-altitude adaptation at the Himalayas. *Molecular Biology and Evolution*, *28*, 1075–1081. doi:10.1093/molbev/msq290

Pilpel, Y. (2011). Noise in biological systems: pros, cons, and mechanisms of control. *Methods In Molecular Biology Clifton Nj*, *759*, 407–425. Retrieved from http://www.springerlink.com/index/10.1007/978-1-61779-173-4

Pirkkala, L., Nykänen, P., & Sistonen, L. (2001). Roles of the heat shock transcription factors in regulation of the heat shock response and beyond. *The FASEB Journal : Official Publication of the Federation of American Societies for Experimental Biology*, *15*, 1118–1131. doi:10.1096/fj00-0294rev

Plotnikov, A., Zehorai, E., Procaccia, S., & Seger, R. (2011). The MAPK cascades: Signaling components, nuclear roles and mechanisms of nuclear translocation. *Biochimica et Biophysica Acta - Molecular Cell Research*. doi:10.1016/j.bbamcr.2010.12.012

Powell, J. E., Henders, A. K., McRae, A. F., Kim, J., Hemani, G., Martin, N. G., … Visscher, P. M. (2013). Congruence of Additive and Non-Additive Effects on Gene Expression Estimated from Pedigree and SNP Data. *PLoS Genetics*, *9*. doi:10.1371/journal.pgen.1003502

Price, N. D., Papin, J. A., & Palsson, B. Ø. (2002). Determination of redundancy and systems properties of the metabolic network of Helicobacter pylori using genome-scale extreme pathway analysis. *Genome Research*, *12*(5), 760–769. doi:10.1101/gr.218002.

Queitsch, C., Sangster, T. A., & Lindquist, S. (2002). Hsp90 as a capacitor of phenotypic variation. *Nature*, *417*, 618–624. doi:10.1038/nature749

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., … Hurles, M. E. (2006). Global variation in copy number in the human genome. *Nature*, *444*(7118), 444–454. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2669898&tool=pmcentrez&rendertype=abstract

Rendel, M. D. (2011). Adaptive evolutionary walks require neutral intermediates in RNA fitness landscapes. *Theoretical Population Biology*, *79*, 12–18. doi:10.1016/j.tpb.2010.10.001

Robertson, K. D. (2005). DNA methylation and human disease. *Nature Reviews. Genetics*, *6*, 597–610. doi:10.1038/nrg1655

Romero, I. G., Ruvinsky, I., & Gilad, Y. (2012). Comparative studies of gene expression and the evolution of gene regulation. *Nature Reviews Genetics*. doi:10.1038/nrg3229

Rutherford, S. L., & Lindquist, S. (1998). Hsp90 as a capacitor for morphological evolution. *Nature*, *396*, 336–342. doi:10.1038/24550

Sauer, U. (2006). Metabolic networks in motion: 13C-based flux analysis. *Molecular Systems Biology*, *2*, 62. doi:10.1038/msb4100109

Scally, A., & Durbin, R. (2012). Revising the human mutation rate: implications for understanding human evolution. *Nature Reviews Genetics*. doi:10.1038/nrg3353

Scally, A., Y, D. J., W, H. L., E, J. G., Ian, G., Javier, H., … C, S. P. (2012). Insights into hominid evolution from the gorilla genome sequence. *Nature*, *483*, 169–175. doi:10.1038/nature10842

Segrè, D., Vitkup, D., & Church, G. M. (2002). Analysis of optimality in natural and perturbed metabolic networks. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(23), 15112–7. doi:10.1073/pnas.232349399

Sequencing, T. C., & Consortium, A. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, *437*(7055), 69–87. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16136131

Shastry, B. S. (2002). SNP alleles in human disease and evolution. *Journal of Human Genetics*, *47*, 561–566. doi:10.1007/s100380200086

Shea, B. T. (2005). Shaping Primate Evolution. Form, Function and Behavior. *International Journal of Primatology*. doi:10.1007/s10764-005-8864-8

Sholtis, S. J., & Noonan, J. P. (2010). Gene regulation and the origins of human biological uniqueness. *Trends Genet*, *26*, 110–118. doi:10.1016/j.tig.2009.12.009

Simon, J., Peifer, M., Bender, W., & O'Connor, M. (1990). Regulatory elements of the bithorax complex that control expression along the anterior-posterior axis. *The EMBO Journal*, *9*, 3945–3956.

Smallbone, K., & Simeonidis, E. (2009). Flux balance analysis: a geometric perspective. *Journal of Theoretical Biology*, *258*(2), 311–5. doi:10.1016/j.jtbi.2009.01.027

Smolke, C. D., & Silver, P. a. (2011). Informing biological design by integration of systems and synthetic biology. *Cell*, *144*(6), 855–9. doi:10.1016/j.cell.2011.02.020

Somerville, M. J., Mervis, C. B., Young, E. J., Seo, E.-J., del Campo, M., Bamforth, S., … Osborne, L. R. Severe expressive-language delay related to duplication of the Williams-Beuren locus. , 353 The New England journal of medicine 1694–1701 (2005). doi:10.1056/NEJMoa051962

Sorensen, P. H., Lessnick, S. L., Lopez-Terrada, D., Liu, X. F., Triche, T. J., & Denny, C. T. (1994). A second Ewing's sarcoma translocation, t(21;22), fuses the EWS gene to another ETS-family transcription factor, ERG. *Nature Genetics*, *6*, 146–151. doi:10.1038/ng0294-146

Spitz, F., & Furlong, E. E. M. (2012). Transcription factors: from enhancer binding to developmental control. *Nature Reviews. Genetics*, *13*(9), 613–26. doi:10.1038/nrg3207

Stankunas, K., Berger, J., Ruse, C., Sinclair, D. A., Randazzo, F., & Brock, H. W. (1998). The enhancer of polycomb gene of Drosophila encodes a chromatin protein conserved in yeast and mammals. *Development (Cambridge, England)*, *125*, 4055–4066.

Stefansson, H., Rujescu, D., Cichon, S., Pietiläinen, O. P. H., Ingason, A., Steinberg, S., … Stefansson, K. (2008). Large recurrent microdeletions associated with schizophrenia. *Nature*, *455*, 232–236. doi:10.1038/nature07229

Stewart-Ornstein, J., Weissman, J. S., & El-Samad, H. (2012). Cellular Noise Regulons Underlie Fluctuations in Saccharomyces cerevisiae. *Molecular Cell*, *45*, 483–493. doi:10.1016/j.molcel.2011.11.035

Stoletzki, N., & Eyre-Walker, A. (2007). Synonymous codon usage in Escherichia coli: selection for translational accuracy. *Molecular Biology and Evolution*, *24*(2), 374–81. doi:10.1093/molbev/msl166

Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazley, C., Thorne, N., … Dermitzakis, E. T. (2007). Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, *315*(5813), 848–853. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2665772&tool=pmcentrez&rendertype=abstract

Streelman, J. T., & Kocher, T. D. (2002). Microsatellite variation associated with prolactin expression and growth of salt-challenged tilapia. *Physiological Genomics*, *9*(1), 1–4. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=11948285

Sumedha, Martin, O. C., & Wagner, A. (2007). New structural variation in evolutionary searches of RNA neutral networks. *BioSystems*, *90*, 475–485. doi:10.1016/j.biosystems.2006.11.007

Suzuki, M. M., & Bird, A. (2008). DNA methylation landscapes: provocative insights from epigenomics. *Nature Reviews. Genetics*, *9*, 465–476. doi:10.1038/nrg2341

Taberlay, P. C., & Jones, P. A. (2011). DNA methylation and cancer. *Progress in Drug Research*, *67*, 1–23. doi:10.1007/978-3-7643-8989-5_1

Tautz, D., & Schmid, K. J. (1998). From genes to individuals: developmental genes and the generation of the phenotype. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *353*, 231–240. doi:10.1098/rstb.1998.0205

Tejedor, J. R., & Valcárcel, J. (2010). Gene regulation: Breaking the second genetic code. *Nature*. doi:10.1038/465045a

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., … Stamatoyannopoulos, J. a. (2012). The accessible chromatin landscape of the human genome. *Nature*, *489*(7414), 75–82. doi:10.1038/nature11232

Tirosh, I., Weinberger, A., Carmi, M., & Barkai, N. (2006). A genetic signature of interspecies variations in gene expression. *Nature Genetics*, *38*(7), 830–834. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16783381

Tischfield, J. A., & Shao, C. (2003). Somatic recombination redux. *Nature Genetics*. doi:10.1038/ng0103-5

Toda, T., Cameron, S., Sass, P., Zoller, M., & Wigler, M. (1987). Three different genes in S. cerevisiae encode the catalytic subunits of the cAMP-dependent protein kinase. *Cell*, *50*, 277–287. doi:10.1016/0092-8674(87)90223-6

Tóth-Petróczy, A., & Tawfik, D. S. (2013). Protein insertions and deletions enabled by neutral roaming in sequence space. *Molecular Biology and Evolution*, 1–33. doi:10.1093/molbev/mst003

Varki, A., Geschwind, D. H., & Eichler, E. E. (2008). Explaining human uniqueness: genome interactions with environment, behaviour and culture. *Nature Reviews. Genetics*, *9*, 749–763. doi:10.1038/nrg2428

Verstrepen, K. J., Jansen, A., Lewitter, F., & Fink, G. R. (2005). Intragenic tandem repeats generate functional variability. *Nature Genetics*, *37*(9), 986–90. doi:10.1038/ng1618

Vinces, M. D., Legendre, M., Caldara, M., Hagihara, M., & Verstrepen, K. J. (2009). Unstable tandem repeats in promoters confer transcriptional evolvability. *Science (New York, N.Y.)*, *324*(5931), 1213–6. doi:10.1126/science.1170097

Von Dassow, G., & Odell, G. M. (2002). Design and constraints of the Drosophila segment polarity module: robust spatial patterning emerges from intertwined cell state switches. *The Journal of Experimental Zoology*, *294*(3), 179–215. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12362429

Waddington, C. H. (1953). Genetic Assimilation of an Acquired Character. *Evolution*, *7*, 118–126. doi:10.2307/2405747

Waddington, C. H. (1956). Genetic Assimilation of the Bithorax Phenotype. *Evolution*, *10*, 1–13. doi:10.2307/2406091

Waddington, C. H. (1959). Canalization of development and genetic assimilation of acquired characters. *Nature*, *183*, 1654–1655. doi:10.1038/1831654a0

Wagner, A. (2005a). Distributed robustness versus redundancy as causes of mutational robustness. *BioEssays*. doi:10.1002/bies.20170

Wagner, A. (2005). *Robustness and Evolvability in Living Systems*. *Evolution* (Vol. 439, p. 367). Princeton University Press. Retrieved from http://books.google.com/books?id=7O1bGwAACAAJ&printsec=frontcover

Wagner, A. (2005b). Robustness, evolvability, and neutrality. *FEBS Letters*, *579*(8), 1772–1778. doi:10.1016/j.febslet.2005.01.063

Wagner, A. (2008). Neutralism and selectionism: a network-based reconciliation. *Nature Reviews. Genetics*. doi:10.1038/nrg2473

Wagner, G. P., Booth, G., & BagheriChaichian, H. (1997). A population genetic theory of canalization. *Evolution*, *51*, 329–347. doi:10.2307/2411105

Wang, R. T., Ahn, S., Park, C. C., Khan, A. H., Lange, K., & Smith, D. J. (2011). Effects of genome-wide copy number variation on expression in mammalian cells. *BMC Genomics*. doi:10.1186/1471-2164-12-562

Wang, Z., & Moult, J. (2001). SNPs, protein structure, and disease. *Human Mutation*, *17*, 263–270. doi:10.1002/humu.22

Warren, A. J. (2002). Eukaryotic transcription factors. *Current Opinion in Structural Biology*. doi:10.1016/S0959-440X(02)00296-8

Whitehouse, I., Flaus, A., Cairns, B. R., White, M. F., Workman, J. L., & Owen-Hughes, T. (1999). Nucleosome mobilization catalysed by the yeast SWI/SNF complex. *Nature*, *400*, 784–787. doi:10.1038/23506

Wilke, C. O., & Drummond, D. A. (2006). Population Genetics of Translational Robustness. *Genetics*, *173*(1), 473–481. Retrieved from http://arxiv.org/abs/q-bio/0509031

Woerner, S. M., Benner, A., Sutter, C., Schiller, M., Yuan, Y. P., Keller, G., … Gebert, J. F. (2003). Pathogenesis of DNA repair-deficient cancers: a statistical meta-analysis of putative Real Common Target genes. *Oncogene*, *22*, 2226–2235. doi:10.1038/sj.onc.1206421

Woese, C. R. (1965). On the evolution of the genetic code. *Proceedings of the National Academy of Sciences of the United States of America*, *54*, 1546–1552.

Woo, Y. H., & Li, W.-H. (2012). Evolutionary Conservation of Histone Modifications in Mammals. *Molecular Biology and Evolution*, *29*(7), 1–11. doi:10.1093/molbev/mss022

Wood, C. R., Boss M A, Patel, P. T., & Emtage, J. S. (1984). The influence of messenger RNA secondary structure on expression of an immunoglobulin heavy chain in Escherichia coli. *Nucleic Acids Research*, *12*(9), 3937–3950.

Wray, G. a, Hahn, M. W., Abouheif, E., Balhoff, J. P., Pizer, M., Rockman, M. V, & Romano, L. a. (2003). The evolution of transcriptional regulation in eukaryotes. *Molecular Biology and Evolution*, *20*(9), 1377–419. doi:10.1093/molbev/msg140

Wray, G. A., Hahn, M. W., Abouheif, E., Balhoff, J. P., Pizer, M., Rockman, M. V, & Romano, L. A. (2003). The evolution of transcriptional regulation in eukaryotes. *Molecular Biology and Evolution*, *20*(9), 1377–1419. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12777501

Wright, S. I., Yau, C. B. K., Looseley, M., & Meyers, B. C. (2004). Effects of gene expression on molecular evolution in Arabidopsis thaliana and Arabidopsis lyrata. *Molecular Biology and Evolution*, *21*(9), 1719–1726. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15201397

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., … Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, *42*, 565–569.

Yang, Z. (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Molecular Biology and Evolution*, *15*, 568–573. doi:10.1093/oxfordjournals.molbev.a025957

Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z. X. P., Pool, J. E., … Wang, J. (2010). Sequencing of 50 human exomes reveals adaptation to high altitude. *Science (New York, N.Y.)*, *329*, 75–78. doi:10.1126/science.1190371

Zhang, J. (2003). Evolution by gene duplication: an update. *Trends Ecol Evol*, *18*, 292–298. doi:10.1016/S0169-5347(03)00033-8

Zhou, J., Lemos, B., Dopman, E. B., & Hartl, D. L. (2011). Copy-number variation: The balance between gene dosage and expression in Drosophila melanogaster. *Genome Biology and Evolution*, *3*, 1014–1024. doi:10.1093/gbe/evr023

Zienolddiny, S., Ryberg, D., Gazdar, A. F., & Haugen, A. (1999). DNA mismatch binding in human lung tumor cell lines, *47*, 15–25.

2. Design Constraints on a Synthetic Metabolism

Tugce Bilgin [1, 2], Andreas Wagner [1, 2, 3]

[1] Institute of Evolutionary Biology and Environmental Sciences, Building Y27-J-54, University of Zurich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland, [2] The Swiss Institute of Bioinformatics, Zurich, Switzerland, [3] The Santa Fe Institute, Santa Fe, New Mexico, United States of America

This chapter was published in PLoS ONE 7(6): e39903.

doi:10.1371/journal.pone.0039903

2.1. Abstract

A metabolism is a complex network of chemical reactions that converts sources of energy and chemical elements into biomass and other molecules. To design a metabolism from scratch and to implement it in a synthetic genome is almost within technological reach. Ideally, a synthetic metabolism should be able to synthesize a desired spectrum of molecules at a high rate, from multiple different nutrients, while using few chemical reactions, and produce little or no waste. Not all of these properties are achievable simultaneously. We here use a recently developed technique to create random metabolic networks with pre-specified properties to quantify trade-offs between these and other properties. We find that for every additional molecule to be synthesized a network needs on average three additional reactions. For every additional carbon source to be utilized, it needs on average two additional reactions. Networks able to synthesize 20 biomass molecules from each of 20 alternative sole carbon sources need to have at least 260 reactions. This number increases to 518 reactions for networks that can synthesize more than 60 molecules from each of 80 carbon sources. The maximally achievable rate of biosynthesis decreases by approximately 5 percent for every additional molecule to be synthesized. Biochemically related molecules can be synthesized at higher rates, because their synthesis produces less waste. Overall, the variables we study can explain 87 percent of variation in network size and 84 percent of the variation in synthesis rate. The constraints we identify prescribe broad boundary conditions that can help to guide synthetic metabolism design.

2.2. Introduction

Among the most important goals of synthetic biology and biotechnology is to engineer organisms with novel properties (Purnick and Weiss 2009; Smolke and Silver 2011). Most current efforts focus on designing subsystems of organisms, such as regulatory circuits (Purnick and Weiss 2009; Sprinzak and Elowitz 2005) or metabolic pathways (Benner and Sismour 2005; Yarmush and Banta 2003). However, recent advances in genomics and genome synthesis have allowed synthetic biology to make great strides towards the ultimate goal of designing new organisms from scratch (Gibson et al. 2010; Gibson et al. 2008; Murtas 2007; Rasmussen et al. 2004; Smith et al. 2003).

To be able to synthesize new life, one needs to understand life's minimal needs. Considerable effort has thus focused on understanding and creating minimal organisms (Forster and Church 2006; Glass et al. 2006; Kuwahara et al. 2007; Mira et al. 2001; Murtas 2007; Nishikawa et al. 2008; Rasmussen et al. 2004). One line of research studies organisms with very small genomes that comprise only a few hundred genes, such as the gamma proteobacteria (Kuwahara et al. 2007), *Blochannia floridanus* (Gil et al. 2003) and *Carsonella ruddii* (Nakabachi et al. 2008). Such organisms are typically endosymbionts or endoparasites, and receive substantial resources from their host (Mira et al. 2001). Although valuable knowledge has been gained by studying these organisms, this property renders them of limited use in understanding minimal requirements for a *free-living* organism.

A second, complementary line of research starts from a complex genome, successfully eliminates genes from it without affecting viability, and thus creates a genome that is (close to) minimal. This is possible because free-living organisms have many genes that are dispensable in any one environment (Forster and Church 2006; Glass et al. 2006; Mizoguchi et al. 2008; Murtas 2007; Rasmussen et al. 2004). Such systematic gene deletion efforts not only provide insight into minimal genomes, they can also help to eliminate the synthesis of undesired molecules, avoid excessive waste production, and thus increase the efficiency with which an organism synthesizes desired molecules (Rude and Schirmer 2009). Based on both computational and experimental approaches of genome reduction, several proposals for the gene complements of minimal organisms have been made. They range in size from 100 genes to more than 300 genes in organisms such as Mycoplasmas (Forster and Church 2006; Glass et al. 2006; Murtas 2007)

One indispensable feature of any living organisms is its metabolism. A metabolism is a complex network of chemical reactions, catalyzed by enzymes that are encoded in genes. It uses sources of energy and chemical elements – nutrients – to synthesize molecules that an organism needs, including precursors of biomass and various secondary products, such as molecules for defense and communication (Kuwahara et al. 2007). The manipulation of metabolism for technological purposes is known as metabolic engineering (Antoni et al. 2007; Bailey 1991; Lee et al. 2008; Rude and Schirmer 2009).

Metabolic engineering has multiple applications. They include the large-scale, fast, and efficient synthesis of pharmaceuticals, chemical reagents, and biofuels (Keasling

2010; Rude and Schirmer 2009; Schirmer et al. 2010; Smolke and Silver 2011; Steen et al. 2010; Steen et al. 2008). The latter class of molecules is especially important given their importance in energy security and in the reduction of greenhouse gas emissions (Antoni et al. 2007; Lee et al. 2008; Mukhopadhyay et al. 2008; Schmidt 2010). Another application is bioremediation, where microbes with properly engineered metabolic pathways may be able to clean hazardous waste in inaccessible places (Cases and Lorenzo 2005; Schmidt 2010).

Current metabolic engineering approaches typically manipulate one or a few enzyme-coding genes (Purnick and Weiss 2009). Because of the highly interconnected nature of metabolism, and because of the complexity of enzyme regulation, such manipulation faces several challenges. The first is to ensure a high level of expression of the genes and the enzymes they encode. A second challenge is to manipulate cells into selectively producing desired molecules at high rates and yield. Cells can be quite recalcitrant to such manipulations (Stephanopoulos and Vallino 1991). A third challenge is to ensure that a desired product can be produced from one source of chemical elements and energy, but from multiple sources, to ensure efficient production. For example, yeast species are good candidate organisms to synthesize ethanol, with the drawback that they are not highly efficient at fermenting cellulosic biomass. (In addition to glucose, which yeast strains can catabolize normally, cellulosic biomass contains five carbon sugars, such as arabinose and xylose, which yeast strains cannot catabolize.) Metabolic engineering can create yeast strains that ferment not only glucose but also mixtures of other sugars. (Wisselink et al. 2009). A fourth challenge is to overcome the toxicity of some desired products when they accumulate at high concentrations in a cell. This holds especially for biofuels that are

produced in large amounts (Mukhopadhyay et al. 2008). Finally, metabolic engineering needs to control ratios of metabolites such as ATP/ADP or NAD+/NADH, which can influence product yields and lead to synthesis of undesired byproducts through their global effects on physiology (Lee et al. 2008).

While contemporary metabolic engineering focuses on altering existing pathways, future engineering will design metabolisms and minimal organisms de novo (Lee et al. 2008; Liang et al. 2011; Purnick and Weiss 2009; Savage et al. 2008). Ongoing technological advances in sequencing and de-novo synthesis and declining prices in these technologies (Antoni et al. 2007; Carr and Church 2009; Shendure et al. 2004; Smolke and Silver 2011) suggest that de-novo synthesis of minimal organisms for biomass production will be feasible soon. A small (synthetic) metabolism may also allow better control of metabolic properties than a large metabolism (Mizoguchi et al. 2008; Purnick and Weiss 2009).

To be able to design a metabolism, one needs to be able to predict system-wide metabolic properties. In recent years great strides have been made towards such prediction. Especially noteworthy are constraint-based modeling approaches, which can predict the spectrum of biosynthetic properties of a metabolism from knowledge about the reactions that its enzymes catalyze, and from the nutrients available in its environment (Edwards and Palsson 2000b; Edwards and Palsson 2000a; Segrè et al. 2002). One such approach, flux balance analysis (FBA) (Edwards and Palsson 2000b; Edwards and Palsson 2000a; Orth et al. 2010; Smallbone and Simeonidis 2009), uses information about the stoichiometry of reactions in a metabolic network to predict the rate at which a network can synthesize a given spectrum of molecules, which we refer

to as the network's biosynthetic flux. FBA makes two main assumptions. The first is that a metabolism is in a steady state with a constant nutrient supply. The second is that it maximizes some property, such as biosynthetic flux (Feist and Palsson 2010). While FBA faces challenges caused by regulatory constraints (Fong and Joyce 2005; Fong et al. 2003), it is well suited to answer simple qualitative and quantitative questions about important properties of a metabolic system (Schellenberger et al. 2007). An especially important property is the minimal number of reactions *R* needed to synthesize a given number *B* of (biomass) molecules from a given spectrum *N* of nutrients. The ideal network has few reactions and can synthesize many molecules using a broad spectrum of nutrients. However trade-offs between these properties exist, which do not allow all these requirements to be met.

We here take a first step towards a quantitative understanding of these and other trade-offs using constraint-based methods. To this end, we study the properties of not just one metabolic network, but of multiple networks that differ in these properties. Experimental techniques are not yet suitable to do that, but computational approaches are (Oberhardt et al. 2009). The approach we use starts with the observation that any one metabolic network exists in a vast space of possible metabolic networks. The approach uses recently developed techniques (Matias Rodrigues and Wagner 2009; Samal et al. 2010) to create unbiased arbitrary large samples of networks from this space, where each network of the sample has a specific property, such as a given number of reactions, a given number of carbon sources it can use, and a given set of molecules that it can synthesize. The underlying sampling technique, Markov Chain Monte Carlo sampling (Paul G. Higgs 2005; Robert and Casella 2004) is a widely used approach with a well-developed statistical theory (Brooks 1998; Ciliberti et al.

2007; Diaconis 2008; Li 1992; Neal 1993). We use it to quantify the trade-offs and thus the design constraints imposed by important metabolic network properties. Specifically, a first part of our analysis focuses on three main properties. The first is nutrient flexibility *N*, that is, the number of different carbon sources a metabolic network can utilize as *sole* carbon sources. The second is a network's biosynthetic ability *B*, that is, the number of biomass molecules that it can synthesize. The third is the number of reactions *R* in a network. We then extend our analysis to further properties, such as the biosynthetic flux *S*, the rate at which biomass molecules are synthesized, and the amount *W* of waste produced.

2.3. Results

Each additional nutrient requires on average two additional reactions.

Due to the presence of alternative metabolic routes that connect many pairs of molecules, most metabolic systems are able to tolerate genotypic changes, such as the deletion of enzyme coding genes. For instance, 80% of single gene deletions in budding yeast have no detectable phenotypic effect in standard laboratory environments (Hillenmeyer et al. 2008). More specifically, metabolic systems are to some extent robust against deletions of enzyme coding genes, because the resulting elimination of metabolic reactions from a metabolic network does not necessarily affect cell viability (Beller et al. 2010; Forster and Church 2006; Glass et al. 2006; Murtas 2007; Schirmer et al. 2010; Steen et al. 2010; Steen et al. 2008).

We first wanted to explore how the minimally needed number of reactions in a network depends on the network's nutrient flexibility and its biosynthetic abilities. For example, a network that can synthesize biomass on an increasing number of different carbon sources will require more metabolic reactions. But how many more? To answer this question we created random *minimal* networks (see Methods) with a given nutrient flexibility *N*, that is, networks that can use *N* sole carbon sources to synthesize all biomass components, but in which every single reaction is essential, such that no reaction can be removed without abolishing viability of the network on at least one of the carbon sources. More specifically, we created 50 random viable networks that can use *N*=20, 40, 60, and 80 carbon sources as sole carbon sources (for a total of 200 networks), and that could synthesize all *B*=63 *E. coli* biomass components. We then studied the relationship between *N* and the number of reactions in these minimal networks (Figure 1a, orange data points). Linear regression analysis showed that the number of needed reactions increases by a number that is statistically indistinguishable from two for every additional carbon source that a network is required to be viable on. Specifically, we found *R* to depend on *N* as $R=(1.9\pm0.1)N + 234$, where the number 0.1 indicates the 95 percent confidence interval of the regression coefficient (see Supporting Information section 'Confidence Interval Calculation' for details). We then asked whether the same relationship between *N* and *R* also holds if we vary biosynthetic ability. To this end, we repeated the analysis just described, but for networks that are able to synthesize *B*=20, B=30, and B=40 biomass components. The slope of the regression line was indistinguishable from that of *B*=63 in all three cases (Figure 1a, green, purple, and blue data; see Figure S3 for distributions of R). In other words, regardless of a network's biosynthetic abilities,

every additional carbon source that a network is required to be viable on requires on average two additional metabolic reactions.

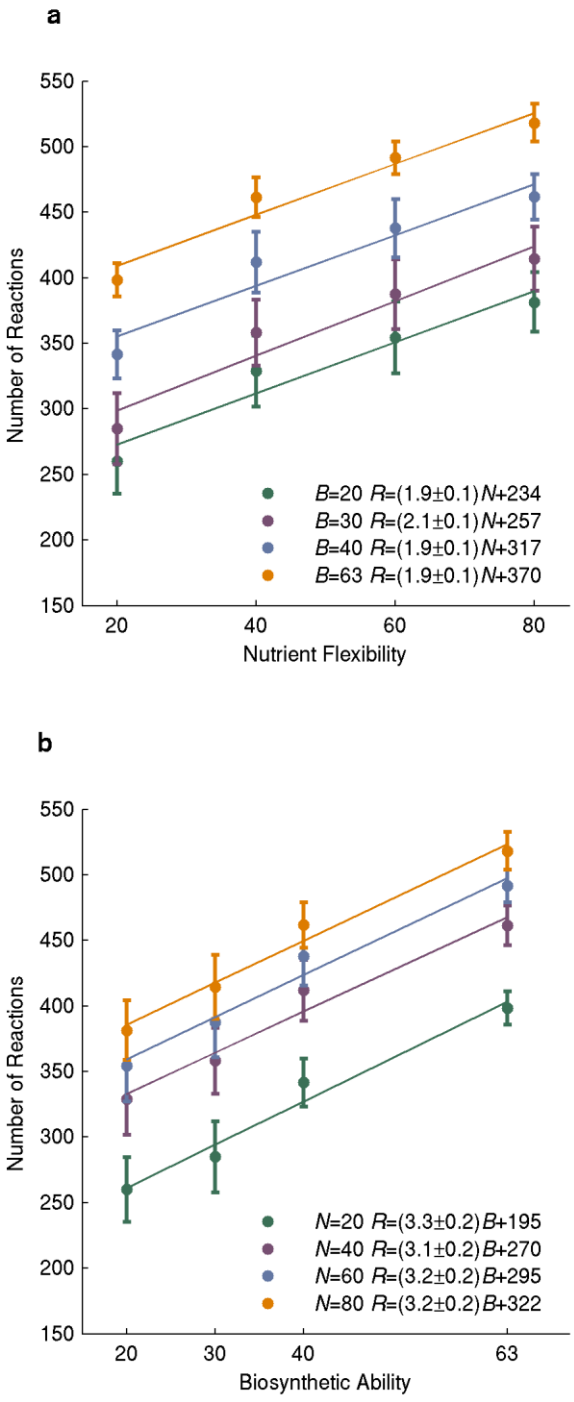

Figure 1. The number of required reactions increases with nutrient flexibility and biosynthetic ability. The vertical axis shows the number of reactions in minimal networks as a function of a) nutrient flexibility and b) biosynthetic ability. Dots and length of error bars correspond to means and one standard deviation based on a sample of n=50 minimal networks. Solid lines indicate linear regression lines for different values of *B* in a) and *N* in b). Numerical estimates of regression coefficients with 95% confidence intervals are given in the inset, in the format $y=(a\pm e)x+b$, where a is the regression coefficient, e the confidence interval, and b is the intercept of the regression line with the vertical axis.

Each additional biomass molecule requires three additional reactions on average.

We next analyzed the relationship between biosynthetic ability and the numbers of reactions in greater detail. To this end, we used the same set of random viable minimal networks that we used in the previous analysis. First, we analyzed the number of reactions in 50 random viable minimal networks with the ability to synthesize B=20, 30, 40, or 63 randomly chosen biomass components, that is, we analyzed a total of 200 networks. Each of these networks was required to be viable on N=80 different sole carbon sources. We found that the number of reactions needed for viability under these conditions increased approximately linearly with biosynthetic ability (Figure 1b, orange), such that every additional biomass molecule required approximately 3 additional reactions Specifically, $R = (3.3 \pm 0.2)B + 195$. We then asked again how this relationship between biosynthetic ability and number of reactions depends on N, and thus repeated this analysis for networks viable on N=20, N=40, and N=60 sole carbon sources. The slope of the regression line was indistinguishable for the different values of N (Figure 1b, green, purple, and blue data), but the intercept differed, as one might expect from the analysis of the previous section.

The results of pairwise regression analysis are easy to display graphically and to interpret intuitively, which is why we use it here. However, where more than two quantities vary, pairwise regression analysis is insufficient to study dependencies among all of them. We thus also carried out a multiple regression analysis using all 800 minimal networks, where nutrient flexibility and biosynthetic ability were independent variables, and where the reaction number was the dependent variable.

Not unexpectedly, the pairwise regression coefficients estimated from the multiple regression analysis were statistically indistinguishable from those estimated from the pairwise analysis above (R=2.0N+3.2B+171). Overall, variation in the two independent variables explains 87 percent of the variation in the number of reactions of a metabolic network.

Our analysis thus far was only concerned with the qualitative question whether networks can synthesize a given number B of molecules, not with the rates at which these molecules can be synthesized. One would therefore expect that our observations are not sensitive to variation in the proportions in which biomass molecules are to be synthesized. An additional analysis (Figure S1) confirms this expectation.

Network size does not significantly influence biosynthetic flux.

Engineered metabolic networks that produce one or more desired products at a high rate are a key goal of biotechnology. For our purpose, it is therefore important to ask whether the maximal rate at which biomass can be synthesized by a network depends on B and N. We here used the ability of FBA to predict the maximal biosynthetic flux (S) of viable networks to answer this question (see Methods).

Our analysis thus far was based on minimal networks, but it is possible that biosynthetic flux depends on the number of reactions in networks that are larger than minimal networks. We therefore first analyzed 1.6×10^4 random viable networks that are not minimal and that differ in size between R=400 and R=800 reactions. (From here on, we will use the term random networks to refer to those networks that are not minimal, unless stated otherwise.) We found that network size does not influence

biosynthetic flux for networks able to synthesize all B=63 *E. coli* biomass molecules. That is, the regression coefficient describing their relationship is statistically indistinguishable from zero (f=rR, with r=$3\times10^{-5}\pm7\times10^{-5}$, n=1000 networks). The same holds for networks that can synthesize B=20, B=30, and B=40 biomass molecules. (See Figure S2 for flux distributions.)

High biosynthetic flux is usually achieved by a modest number of active reactions.

We just showed that the number of reactions in a network does not affect its biosynthetic flux. However, it is well-known that only a subset of a network's metabolic reactions are usually *active* in any one environment, that is, they have nonzero flux (Nishikawa et al. 2008). We thus wanted to know whether a relationship exists between biosynthetic flux and the number of active reactions, R_A. This is indeed the case, based on an analysis of all our 1.6×10^4 random networks viable on glucose as a sole carbon source. More specifically, this relationship is strongly negative (Pearson's r = -0.65 for biosynthetic flux on glucose). That is, the greater the number of active reactions, the lower the biosynthetic flux of a network.

We next used linear regression to ask whether network size R itself influences the number of active reactions R_A. The answer is no. Network size has no statistically significant influence on the number of active reactions for any combination of values of N and B. This observation is in agreement with earlier results by Nishikawa *et al.* (Nishikawa et al. 2008), which showed that the number of active reactions in a network is not sensitive to the size of the network.

We subsequently asked whether synthesis of each additional biomass molecule also needs more active reactions. The answer is yes, as shown by linear regression analysis (3.6-4 additional reactions, on average, per biomass molecule).

The more biomass molecules a network synthesizes, the smaller is its biosynthetic flux.

We next wanted to explore why the number of active reactions is negatively correlated with biosynthetic flux. Our observations so far show that the number of active reactions increases with biosynthetic ability. We thus hypothesized that increased biosynthetic ability entails smaller biosynthetic flux, because biosynthetic flux should decrease as the number of biomass molecules to be synthesized grows, given a constant nutrient supply. To test this hypothesis, we computed the correlation between the number of molecules synthesized and biosynthetic flux in multiple environments for all 16000 networks we had generated (see Methods). This correlation is significantly negative (Pearson's r = -0.60; $P<10^{-300}$). Each molecule to be synthesized decreases biosynthetic flux by 0.05 mmoles per g DW per hour, equivalent to a 6 percent decline relative to *E.coli*'s computed biosynthetic flux. Next, we analyzed networks viable on 80 different sole carbon sources that were able to synthesize B=20, 30, 40, or 63 randomly chosen biomass components (1000 networks each, for a total of 4000 networks). Biosynthetic flux under these conditions also decreased approximately linearly with biosynthetic ability (Figure 2a, orange data). Specifically, every additional biomass molecule that a network needs to synthesize decreases the biosynthetic flux by approximately 0.05 mmoles of biomass per g DW per hour (6 percent; S=-(0.05±0.001)B + 3.5). Similar relationships also hold for networks viable on N=20, N=40, and N=60 sole carbon sources (Figure 2a, green,

purple, and blue data). They show that, every additional biomass molecule to be synthesized reduces biosynthetic flux by 3-6 percent. Taken together, these observations help explain the negative relation between the number of active reactions and biosynthetic flux. Networks that can synthesize more biomass molecules need more active reactions. Given a constant nutrient supply, the total biosynthetic flux that can be realized by these networks must decrease as the number of molecules that they can synthesize increases. The negative correlation between the number of active reactions and biosynthetic flux is a by-product of the latter two correlations.

Biosynthetic flux increases weakly with nutrient flexibility.

We next explored how the nutrient flexibility of a network affects its biosynthetic flux. To this end, we analyzed 1000 networks each that can use N=20, 40, 60, and 80 carbon sources (a total of 4000 networks) and that can synthesize B=63 biomass molecules. (Figure 2b, orange data points). We found that biosynthetic flux increases weakly but significantly with nutrient flexibility (S= (0.01±0.0001)N + 0.6 mmoles of biomass per g DW hour, n=1000 networks) for every additional carbon source that a network is viable on. This is equivalent to 1 percent of increase relative to *E. coli*'s biosynthetic flux on glucose. Quantitatively similar relationships hold for networks able to synthesize B=20, B=30, and B=40 biomass components (Figure 2b, green, purple, and blue data). We next turn to a variable that can help explain this association: waste production.

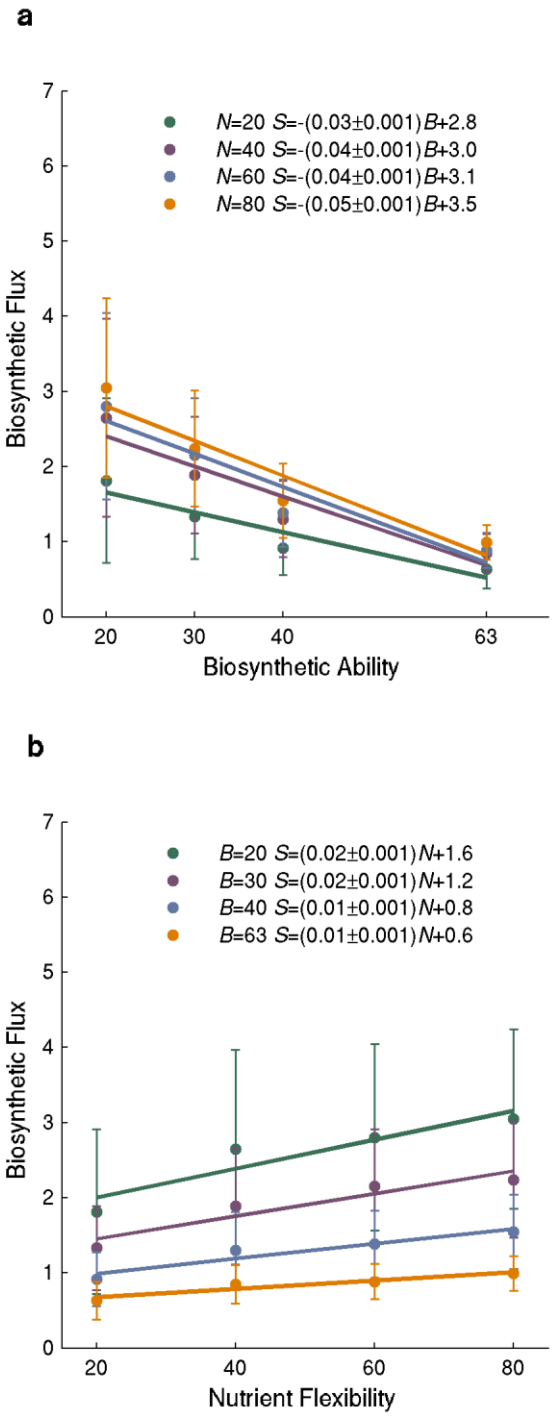

Figure 2. Biosynthetic flux decreases with biosynthetic ability and increases with nutrient flexibility. The vertical axis shows biosynthetic flux in mmoles per g DW per hour in random networks as a function of a) biosynthetic ability and b) nutrient flexibility. Dots and lengths of error bars correspond to means and one standard deviations based on a sample of n=1000 minimal networks. Solid lines indicate linear regression lines for different values of *B*. Numerical estimates of regression coefficients with 95% confidence intervals are given in the inset, in the format $y=(a\pm e)x+b$, where a is the regression coefficient, e the confidence interval, and b is the intercept of the regression line with the vertical axis.

High biosynthetic flux means less waste.

Some 300 transport reactions are described in *E. coli* (Feist et al. 2007), many of which transport waste products either from the periplasm or the cytoplasm to the

extracellular space. Metabolic networks may differ in the extent to which they produce waste products that are excreted from cells. Waste products may include molecules that are not biomass molecules (such as carbon dioxide or acetate), as well as excess biomass molecules (Feist et al. 2007; Nishikawa et al. 2008). Waste production consumes resources, such as organic carbon, and will therefore reduce biosynthetic flux. With these considerations in mind, we next asked whether lower waste production might be responsible for the differences in biosynthetic flux we observed in networks with different nutrient flexibility.

Different metabolic networks may excrete different kinds of molecules, but to compare their waste production, it is useful to establish a common waste 'currency'. Since our analysis is focused on carbon metabolism, we use the moles of carbon per g DW per hour as our unit of waste production. (Note that these moles of carbon may well come from a broad spectrum of different molecules.)

We first wanted to know if waste has any influence on biosynthetic flux. The answer is yes, based on an analysis of all our 1.6×10^4 random viable networks on glucose as the sole carbon source. This relationship is strongly negative (Pearson's r=-0.66) That is, the greater waste production is, the lower the biosynthetic flux.

In the last section, we showed that for each additional carbon source a network was required to be viable on, biosynthetic flux increased by 0.01-0.02 mmoles of biomass per g DW and hour (Figure 3). If we express biomass synthesis instead as moles of carbon (in synthesized biomass) per g DW and per hour, biosynthetic flux increases on average by 0.1 mmoles of carbon per g DW hour with increasing nutrient

flexibility. We hypothesized that this increase can be explained through the influence of nutrient flexibility *N* on waste production *W*. To this end, we first analyzed random viable networks that vary in *N* and that can synthesize all *B*=63 molecules *E. coli* biomass molecules (Figure 3, orange data points). We found that waste production *decreased* by a number that is statistically indistinguishable from 0.1 mmol of carbon per g DW hour ($W=(-0.09\pm0.02)N+24$, n=1000 networks) for every additional carbon source that a network is viable on. Statistically indistinguishable relationships exist for networks with *B*=20, *B*=30, and *B*=40 biomass components. (Figure 3, green, purple, and blue data). These observations suggest that the positive influence of nutrient flexibility on biosynthesis flux comes from reduced waste production. We later discuss experimental evidence supporting this observation.

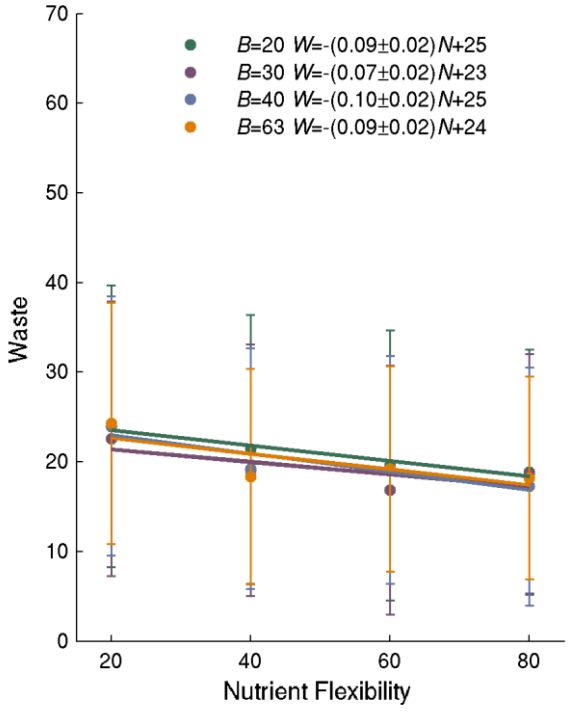

Figure 3. Waste production decreases with nutrient flexibility. The vertical axis shows waste that is excreted carbon in mmoles per g DW hour in random networks as a function of nutrient flexibility. Dots and lengths of error bars correspond to means and one standard deviation based on a sample of n=1000 minimal networks. Solid lines indicate linear regression lines for different values of *B*. Numerical estimates of regression coefficients with 95% confidence intervals are given in the inset.

In a final analysis related to waste production, we studied the relationship between biosynthetic ability and waste production, and found no significant such relationship.

The variables we considered can explain 84 percent of the variance in biosynthetic flux.

Thus far, we have considered five variables and how they influence biosynthetic flux. These are network size R, nutrient flexibility N, biosynthetic ability B, waste production W, and numbers of active reaction R_A. To understand how biosynthetic flux depends on not just one but all of these variables, we carried out a multiple regression analysis with flux as the dependent variable. This regression analysis showed that the variables we analyzed explain 84 percent of the variation in biosynthetic flux (R^2=0.84).

Synthesis of biochemically related molecules requires fewer reactions, because it produces less waste.

In our analysis of how required reaction numbers depend on a network's biosynthetic abilities, we have purposely focused on randomly chosen biomass precursors, as they give us an unbiased view of this dependency. However, this relationship may change if one considers biochemically related biomass molecules. To obtain some insights how it changes, we next studied a group of molecules whose known biosynthesis pathways share several reactions. These are the 20 proteinaceous amino acids (Morot-Gaudry et al. 2001). Figure 4a shows the number of reactions in minimal networks that can synthesize all 20 proteinaceous amino acids (left panel), as well as the number of reactions needed to synthesize 20 randomly chosen biomass molecules (right panel). Amino acid synthesizing networks need 191±10 (mean ± one standard

deviation) reactions, 27 percent fewer than the 261±24 reactions needed to synthesize 20 arbitrary molecules. This difference is highly significant ($P<10^{-10}$, Mann-Whitney U-test, n=100) (Mann and Whitney 1947).

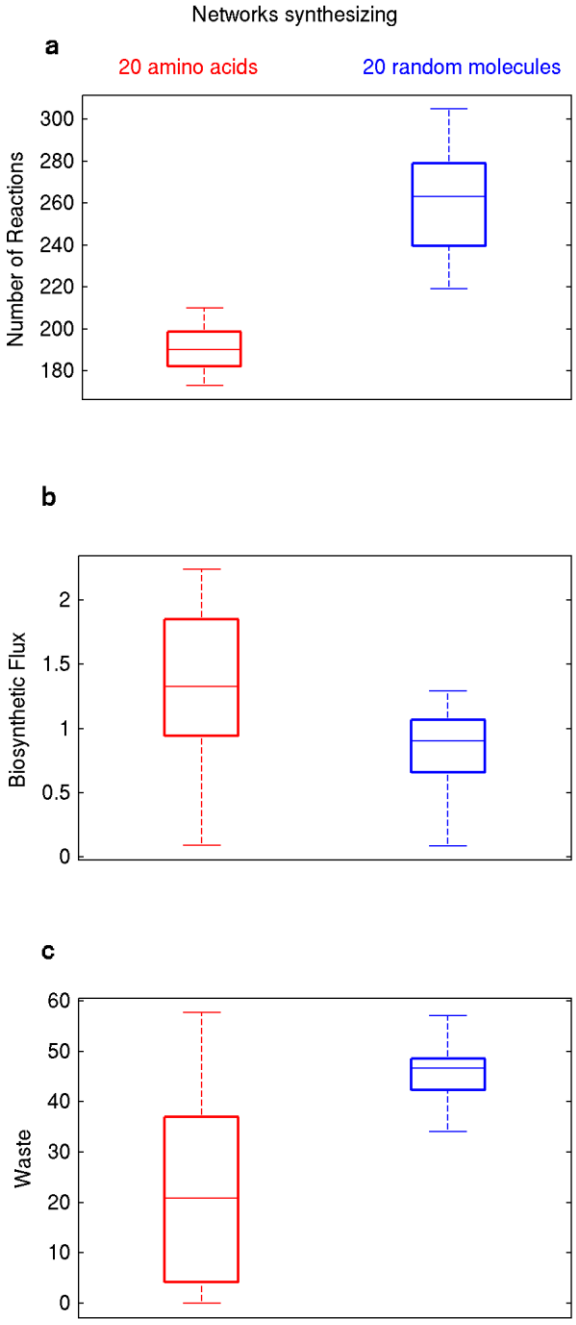

Figure 4. Synthesis of biochemically related compounds require less reactions, achieve higher biosynthetic flux with less waste. Box-plot of a) number of reactions, b) biosynthetic flux in mmoles per g DW hour, and c) waste in mmoles per g DW per hour in minimal networks synthesizing twenty amino acids (left panel) and synthesizing twenty random biomass molecules (right panel). Horizontal lines in the middle of each box mark the median. The edges of the boxes correspond to the 25th and 75th percentiles. Data is based on a sample of *n*=80 for each box.

In minimal networks every reaction is active under the conditions we study. Our observations in the last paragraph thus also show that amino acid biosynthesis needs fewer active reactions than biosynthesis of arbitrary biomass molecules. One of our analyses above showed that fewer active reactions also imply higher biosynthetic flux, which raises the question whether the 20 amino acids can be synthesized at higher rates. The answer is yes (Figure 4b). Minimal networks that synthesize 20 arbitrary biomass molecules synthesize them 35 percent more slowly than minimal networks that synthesize 20 amino acids ($P<10^{-13}$, n=80, Mann-Whitney U-test) (Mann and Whitney 1947).

We next hypothesized that these dependencies might be explicable as a by-product of a lower cost of amino acid synthesis. If amino acids have fewer carbon molecules than the average biomass molecule, both fewer steps would be needed to synthesize them and a higher synthesis rate could be achieved. To test this hypothesis, we computed the average carbon content of all 20 amino acids, weighted by each amino acid's stoichiometry in biomass, as 1.28±0.60 mmoles of carbon per mmole biomass. This molecular weight turned out not to be smaller but significantly larger than the carbon content of 20 random biomass molecules, also weighted by their stoichiometry in biomass (0.87±0.15 mmoles of carbon, $P<2.2\times10^{-16}$, one sample t-test, for n=105 sets of 20 random biomass molecules). This means that low amino acid cost cannot explain a higher rate of amino acid production.

We next asked whether lower waste production might be responsible for the lower biosynthetic flux we observed. We therefore analyzed the quantity of secreted waste products in 80 networks that synthesize 20 amino acids, and in 80 networks that

synthesize 20 random biomass compounds. Waste production is indeed significantly lower in networks that synthesize amino acids ($P<10^{-17}$, Mann-Whitney U-test) (Mann and Whitney 1947). Specifically, amino acid synthesizing networks produce 47 percent less waste than networks that synthesize arbitrary molecules. They excrete on average 22.8 mmoles of carbon waste per g DW hour (Figure 4c).

To provide a concrete example of a prominent waste molecule, consider the amount of secreted carbon dioxide. For each mole of carbon entering a network in the form of glucose, amino acid synthesizing networks excrete 0.18±0.12 (mean ± one standard deviation) moles of carbon per g DW hour as carbon dioxide, whereas networks synthesizing 20 arbitrary molecules excrete 0.31±0.06 mmoles of carbon per g DW hour as carbon dioxide ($P<10^{-25}$, Mann-Whitney U-test) (Mann and Whitney 1947). In sum, waste production is an important factor in explaining the smaller costs of amino acid biosynthesis compared to arbitrary biomass molecules.

2.4. Discussion

We here studied typical properties of metabolic networks with recently developed techniques to create random and unbiased samples of metabolic networks from a vast space of such networks (Matias Rodrigues and Wagner 2009; Samal et al. 2010). Specifically, we studied the quantitative relationships between 6 different properties. These are the number of alternative sole carbon sources *N* that a network can use – its nutrient flexibility –, the number of (biomass) molecules *B* it can synthesize, the rate *S*

at which it can synthesize these molecules, the number of reactions R in the network, the number of active reactions R_A, that is, reactions that have nonzero metabolic flux, and the amount of waste W the network produces.

We focused first on how the number of minimally needed reactions R depends on N and B, because this number of reactions is an important design variable. Smaller networks would be easier to design and smaller genomes would be easier to synthesize (Carr and Church 2009). More than that, biosynthetic processes are more controllable and predictable in a small metabolism (Mizoguchi et al. 2008; Purnick and Weiss 2009). We found that for every additional molecule to be synthesized, a network needs on average three additional reactions.

In a related analysis, we focused on nutrient flexibility. The ability of a metabolic network to sustain life on multiple sources of chemical elements and energy is highly desirable in industrial production processes (Wisselink et al. 2009). Examples include biofuel production from cellulosic biomass. Cellulosic biomass contains both hexoses (e.g., glucose) and pentoses, whose major constituent is xylose (Ho et al. 1998; Wisselink et al. 2009). Many yeast species, for instance, do not consume pentoses, and need to be engineered to have greater nutrient flexibility to make biofuel production more efficient (Ho et al. 1998). In our analysis, we found that for every additional carbon source to be utilized, a network needs on average two additional reactions. Thus, it is more expensive, in terms of the number of needed reactions, to synthesize additional molecules than to synthesize the same molecules but from a larger number of alternative carbon sources.

Anecdotal evidence for the tendency that increased nutrient flexibility requires larger networks is provided by two networks for which nutrient flexibility is experimentally known (Feist et al. 2007; Oh et al. 2007). These are *E. coli* where N=54 and R=1396 and *Bacillus Subtilis* where N is smaller (N=41) and so is R (R=769). We also compared the computationally predicted number of carbon sources that can be utilized by metabolic models of seven organisms (Column 4 of Table 1). Although based on few data points, this analysis suggests that there is a significant correlation (Pearson's r=0.9, P=0.005) between the *in silico* nutrient flexibilities and networks sizes of these model organisms.

Organism	***B***	***R***	R_A	***N***	***References***
Buchnera Aphidicola	43	205	183	2	(Thomas et al. 2009)
Helicobacter pylori	52	394	298	35	(Schilling et al. 2002)
Staphlococcus aureaus	58	534	286	41	(Becker and Palsson 2005)
Bacillus Subtilis	59	769	327	97	(Oh et al. 2007)
Methanosarcina barkeri	63	531	352	6	(Feist et al. 2006)
Escherichia coli	67	1396	388	175	(Feist et al. 2007)
Mycobacterium tubeculosis	93	836	408	39	(Beste et al. 2007)

Table 1. Biosynthetic ability *B*, number *R* of metabolic reactions (size of metabolic networks), number R_A of number of active reactions and number *N* of carbon sources predicted to be catabolized in computational metabolic models of various organisms. We took the number of synthesized molecules *B*, and the number of metabolic reactions *R* from the genome scale metabolic network reconstructions of those organisms as listed in the column *'References'*. Active reactions are reactions that have non-zero flux as determined by FBA. We computed nutrient flexibilities by testing the viability of metabolic models on all known carbon containing metabolites that are educts or products of reactions in the global network by FBA and assumed a metabolite to be a carbon source if the metabolic model can utilize it.

As a result of these dependencies, the minimally needed number of reactions varies widely among the networks we study. For example, for networks that can use 20 alternative carbon sources, it ranges from an average of 260 reactions (which would be, for comparison 19% of the whole *E. coli* network) needed to synthesize *B*=20 molecules to an average of 381 reactions (29% of the *E. coli* network) to synthesize all *B*=63 biomass molecules of *E. coli*. For networks that can use 80 alternative carbon sources, the largest number we studied, it increases from 398 (27%) to 518 reactions (37% of the *E. coli* network) as *B* increases from 20 to 63.

N and *B* can explain the vast majority (87%) of the variance in *R*, in a relationship that is close to linear. Note that the proportions in which biomass molecules are to be synthesized are immaterial, as long as the metabolic reactions to synthesize each required molecule are present. This is why the stoichiometry of biomass composition would not influence the relationships we observe. The remaining 13% is variance unexplained by linear relationships among the variables we considered. Such unexplained variance could have multiple sources, and especially nonlinear relationships among the variables. For example, as the number of biomass molecules to be synthesized increases, fewer and fewer additional reactions might be required, because most precursors of the additional biomass molecules may already be synthesized by existing reactions.

We showed that the reactions that are needed to utilize an additional carbon-containing molecule or to synthesize an additional biomass molecule typically involve the additional molecule. For example, the utilization of an additional carbon source requires a reaction that catabolizes that molecule. The synthesis of an additional

molecule requires a reaction that produces that molecule. We emphasize that the required numbers of reactions we found for an additional carbon source or biomass molecule are statistical patterns, averages over multiple networks. Some additional carbon sources or biomass molecules need more reactions than the average we identified, whereas others need fewer. (For some examples see Supporting Information, Table S3, Table S4, sections 'Examples of reactions needed to metabolize new carbon sources' and 'Examples of reactions needed to synthesize additional biomass molecules'.)

We also studied the number of active reactions – reactions with nonzero flux on a given carbon source. Quantitatively, we found that the number of active reactions increases by 3.6-4 reactions for each additional biomass molecule to be synthesized. Qualitatively, this tendency – increasing synthetic ability requires more active reactions – is not surprising. Table 1 shows, as an example of this tendency, several well-known organisms with known metabolic network sizes R and biosynthetic abilities B. Note that, with exceptions, the number of active reactions R_A tends to become higher as B increases for these organisms. Albeit based on few species, a regression analysis of the data in Table 1 shows that the relationship between R_A and B is similar to what we find in our much larger samples of random networks (R_A=3.4B+120). Our approach has the advantage of allowing us to make quantitative statements about the number of active reactions required to synthesize additional molecules that are not just based on few, well studied organisms, as in Table 1, but on arbitrarily large and random samples of metabolic networks.

The identity of active metabolic reactions can depend strongly on the nutrient environment (Nishikawa et al. 2008; Samal et al. 2010; Wang and Zhang 2009). A reaction active in one environment may be inactive in another. Organisms would cope with such variation by regulating the activity of reactions, for example by regulating the expression of enzyme-coding genes or by regulating the enzymes themselves (Cooper 2008). The importance of reaction activity in our analysis thus points towards the importance of regulating metabolic flux in response to the environment. Engineering optimal regulation of enzymes is a key challenge in metabolic engineering (Kim and Reed 2010; Wessely et al. 2011). It will be an even greater challenge in a synthetic metabolism, especially if such a metabolism is to function *efficiently* in multiple chemical environments.

A second major set of analyses focused on how metabolic network properties influence biosynthetic flux *S*. These analyses are motivated by the fact that high synthesis rates of one or more target molecules are important goals of metabolic engineering (Antoni et al. 2007; Lee et al. 2008; Rude and Schirmer 2009). Our first major observation in this regard concerns the number of reactions in a metabolic network. This number has virtually no influence on the attainable biosynthetic flux *S*. In stark contrast, the number of active reactions – reactions with nonzero flux – shows a strong negative association with biosynthetic flux. That is, the higher the biosynthetic flux is, the smaller is the number of active reactions. (This analysis is based on networks that have more than the minimally necessary number of reactions in a given environment. In minimal networks, all reactions must be active.) This observation is consistent with earlier work based on four microbial metabolic networks (Nishikawa et al. 2008).

A second important influence on biosynthetic flux *S* is the number of biomass molecules to be synthesized. Specifically, every additional such molecule reduces *S* by approximately 5 percent. This influence is stronger than the influence of nutrient flexibility, where each additional carbon source changes biomass synthesis flux by 1 percent. That nutrient flexibility influences biosynthesis rates is known from experiment (Garcia Sanchez et al. 2010; Liu and Hu 2010; Madhavan et al. 2009; Nevoigt 2008; Wisselink et al. 2009). For example, engineered yeast strains capable of growing on more carbon sources achieve higher product yields (Wisselink et al. 2009).

The amount of waste *W* that a network produces is a third, and the most important influence on biosynthetic flux *S*. Metabolic networks generally excrete waste, for example, in the form of carbon dioxide (Nishikawa et al. 2008), acetate (Nishikawa et al. 2008; Weber et al. 2005), pyruvate (Weber et al. 2005) or glycerol (Ho et al. 1998; Nevoigt 2008). Not only does this mean wasted resources, some waste products may also be toxic and interfere with goals of metabolic engineering (Mukhopadhyay et al. 2008). We found that biosynthetic flux *S* shows a strong negative correlation with waste production *W* in the form of excreted carbon. This relation is in agreement with experimental observations (Ho et al. 1998; Nevoigt 2008; Weber et al. 2005). For example in *E. coli*, Weber *et al* (Weber et al. 2005) demonstrated that excretion of methylglyoxal, D- and L-lactate, pyruvate, and acetate decreases growth rates. We also observed that networks with higher nutrient flexibility produce less waste, such that for each additional carbon source, networks produce 0.1 mmoles less in excreted carbon waste. This relationship can help explain the positive influence of *N* on flux *S*. Experimental support for this observation exists as well. For example, Wisselink *et al*.

(Wisselink et al. 2009) observed that increased ethanol production in engineered yeast with higher nutrient flexibility is caused by reduced production of the waste products xylitol and arabinitol. Relatedly, the elimination of glycerol formation increases the yield of ethanol up to 10 percent in yeast (Nevoigt 2008). Ho *et al*. (Ho et al. 1998) report similar observations.

We also found that an analogous association does not exist for *B.* Networks seem to produce indistinguishable amounts of waste products regardless of how many biomass molecules they synthesize.

Overall, the quantities we examined can explain 84 percent of the variance in biosynthetic flux *S* we observed. This means, that these quantities account for most of the variation in the biosynthetic flux, and are thus important factors in the design of a synthetic metabolism.

In our final analysis, we showed that the nature of the molecules to be synthesized can influence biosynthetic flux. For example, in biotechnological applications, a metabolic network may need to synthesize molecules that are closely related, and whose biosynthetic pathways are therefore similar, in the sense that they share many reactions. Examples include vitamins and coenzymes (Vandamme 1992), taxols and related taxoids (Expósito et al. 2009), hydrocarbons and related ether lipids (Metzger and Largeau 2005), amino acids (Leuchtenberger et al. 2005). We examined the influence of biochemical relatedness by studying metabolic networks that synthesize all 20 amino acids, and comparing them with networks that synthesize 20 arbitrary biomass molecules. In this analysis, we found that the amino acids could be

synthesized by networks with fewer reactions, thus demonstrating their biosynthetic relatedness. They can also be synthesized at a rate that is 35 percent higher, because their biosynthesis produces 47 percent less waste as excreted carbon. These differences are substantial, given that metabolic engineering in biotechnological processes typically increases synthesis rates of desired products by 5-10 percent (Dueñas-Sánchez et al. 2010; Nevoigt 2008; Raab et al. 2011; Raghevendran et al. 2006).

We conclude with some caveats and limitations of our analysis. First, we are well aware that there is still a gap between current metabolic engineering and synthetic biology experiments, and theoretical studies such as ours. We see the value of studies like ours as providing quantitative reference points for future experimental work in this area.

Second, flux balance analysis, on which our approach rests, assumes a metabolic system that is in a steady-state, such as might be achieved by a microbial population growing under a constant nutrient supply. This assumption ignores possible additional constraints that regulation and enzyme kinetics exert on a metabolism (Kauffman et al. 2003; Price et al. 2004). While achieving proper regulation remains a big challenge in synthetic biology, we note that some of our observations are unlikely to be affected by this assumption. For example, relaxing this assumption would not reduce the minimal number of reactions needed for a given biosynthetic task.

Third, we focus on *typical* network properties, that is, properties of metabolic networks sampled at random from a large space of such networks. Optimization

procedures that search metabolic networks space systematically may be able to identify individual networks whose properties deviate from those we identified as typical. For example, they may identify networks that are able to synthesize biomass with even fewer reactions or at higher rates. Such procedures have been successful in other combinatorial optimization problems. They include simulated annealing (Gonzalez et al. 2007; Tomshine and Kaznessis 2006), bi-level optimization (Domingues et al. 2010; Yang et al. 2008), OptFlux (Rocha et al. 2010), convex optimization (Julius et al. 2008) and evolutionary optimization (Ebenhöh and Heinrich 2001; Patil et al. 2005). The extent to which they can identify networks with superior design remains an important subject of future work.

Fourth, the expression of enzymes itself has a metabolic cost. This cost may be reduced by regulating enzyme expression depending on the nutrient environment (through regulatory machinery whose expression may itself be costly.) We did not consider costs like these here, because they are poorly understood on a quantitative level. Their analysis remains an important goal for future work.

Finally, we note that an organism's genome encodes more than metabolism. Nonmetabolic genes, even in small genomes serve roles in systems that allow cell motility, signaling, secretion, and defense (Kuwahara et al. 2007). The heterogeneity of these systems will make general, quantitative statements about design constraints for entire genomes more difficult to obtain. In this regard, we note that the proportion of a gene's genome devoted to metabolism increases in small genomes. For example, in the free-living *E. coli* with a large genome, fewer than 75 percent of genes exert metabolic functions (Riley 1997). In contrast, in some endosymbionts including

Buchnera, *Wiggleswothia glossinidia* and *Thiomicrospira crunogena* more than 80 percent of genes are devoted to metabolic functions, and in yet others, such as *the gamma proteobacteria, Blochmanni floridanus* and *Wolbachia pipientis* more than 95 percent have metabolic functions (Kuwahara et al. 2007). Thus, the constraints we identified here would affect the majority and perhaps the vast majority of a synthetic minimal organism's genome.

2.5. Methods

The *metabolic genotype* of an organism comprises all genes that encode metabolic enzymes. This genotype can be compactly represented through a list of reactions that can take place in the organism and that are catalyzed by enzymes (Edwards and Palsson 2000b; Edwards and Palsson 2000a; Goto et al. 2002; Goto et al. 2000; Kanehisa and Goto 2000; Klasson 2004). Each metabolic genotype exists in a vast metabolic *genotype space* that contains all possible metabolic networks – all possible combinations of reactions drawn from a set of biochemically feasible reactions – that *could* take place in a living system. According to our current knowledge, there are more than 5000 such reactions, which means that metabolic genotype space comprises 2^{5000} possible metabolic networks. The metabolic genotype or metabolic network of any one organism can be viewed as a point in this space. Two genotypes differing from each other in a single reaction are *neighbors* in this space. We refer to a network's metabolic *phenotype* as the spectrum of chemical environments – defined by nutrients these environments contain – on which the network can synthesize a

predetermined spectrum of molecules, as well as the rate at which it can synthesize these molecules. We call a network *viable* in a given environment if it can synthesize all these molecules in the environment.

Flux Balance Analysis

Flux balance analysis (FBA) is a constraint-based modeling approach that predicts steady state metabolic fluxes – rates of substrate to product conversion – for all metabolic reactions in a metabolic network. It can thus also predict the biomass yield and other complex metabolic attributes of a metabolic network (Kauffman et al. 2003; Price et al. 2004). Because FBA does not need kinetic information, but only stoichiometric information about the reactions involved, it is a widely used approach for analyzing the metabolism of well-studied organisms such as *E. coli* and *S. cerevisiae* (Kauffman et al. 2003; Orth et al. 2010; Price et al. 2004). The needed stoichiometric information comes from experimental biochemical analysis, as well as from comparative analyses of enzyme-coding genes in different, completely sequenced genomes. This information is encapsulated in a stoichiometric matrix **S** of dimensions $m \times n$, where m denotes the number of metabolites, and n is the number of reactions in a network (Kauffman et al. 2003; Orth et al. 2010). FBA assumes that a metabolic network is in a metabolic steady-state, such as may occur in a microbial population that proliferates in an unchanging environment. Because mass needs to be conserved under these conditions, any vector **v** of metabolic fluxes – the rates at which a network's metabolic reactions convert substrates into products – must satisfy the equation

$$\mathbf{Sv} = 0$$

This equation typically has a large solution space of allowable fluxes. The size of this space can be reduced somewhat by placing biochemically motivated constraints on the irreversibility of some reactions and on maximal flux magnitudes (Covert et al. 2001). In the (reduced) solution space, FBA then uses linear programming to identify regions in the space that maximize a quantity of interest, which can be represented by a linear objective function Z (Kauffman et al. 2003; Orth et al. 2010). More specifically, the linear programming formulation of an FBA problem can be written as:

$$\max Z = \max \{\mathbf{c}^T \mathbf{v} \mid \mathbf{S}\mathbf{v} = 0, \mathbf{a} \leq \mathbf{v} \leq \mathbf{b}\} \quad (1)$$

where the vector $\mathbf{c}^T$ stands for a transposed (^{T}) array **c** of scalar coefficients that define the objective function. Vectors **a** and **b** contain lower and upper limits of reaction fluxes in the flux vector **v**, respectively. A particularly important quantity to be maximized is the rate of synthesis of a given set biomass molecules. We refer to this quantity as the *biosynthetic flux* (in units of mmol per g DW per hour) of a metabolic network. We used the software packages CPLEX (11.0, ILOG; http://www.ilog.com/) and CLP (1.4, Coin-OR; https://projects/coinor.org/Clp) to solve all linear programming problems that arise in this study.

Biomass Molecules

The starting point of our analysis was a set of 67 biomass molecules from *E. coli* (Feist et al. 2007). We used these molecules, because *E. coli's* biomass composition is well studied, and many of its components – amino acids, nucleotides, etc. – also occur in most other free-living organisms. In addition, *E. coli* and its biomass molecules are

highly relevant to biotechnological applications (Lee et al. 2008; Mizoguchi et al. 2008; Rude and Schirmer 2009). For some of our analyses, we needed to vary the spectrum of metabolites that a metabolic network needs to synthesize. Among the 67 *E. coli* biomass molecules, we allowed all but four molecules to vary. These four molecules are the "currency metabolites" adenosine 5'-diphosphate (ADP), phosphoric acid (Pi), pyrophosphoric acid (PPi), and hydrogen ions. A complete list of the biomass molecules we used can be found in Supplementary Table S1.

Carbon Sources

Any one metabolic network can only synthesize biomass if its chemical environment contains specific nutrients (Handorf et al. 2008). In FBA, these nutrients are represented by special exchange reactions that reflect transport of nutrients into the cell. The environments we study are minimal chemical environments that contain a single molecular source for each essential chemical element. These sources are oxygen, ammonium, inorganic phosphate, sulfate, sodium, potassium, cobalt, iron, protons, water, molybdate, copper, calcium, chloride, magnesium, manganese and zinc, as well as a single source of carbon. We study how variation in carbon sources constrains the composition of metabolic networks that needs to synthesize a given spectrum of molecules. We chose to vary carbon sources for this purpose, because of carbon's centrality as a chemical element in biomass. A complete list of the carbon sources we used is given in the Supplementary Table S2.

We computed the biomass growth of a network by taking the average of its biomass growth rates on each carbon source that the network is required to be viable on, except where mentioned otherwise.

Global Network

Some of our analyses use a "global reaction network". This global network contains a comprehensive set of known biochemical reactions from multiple organisms, and it has universal biosynthetic abilities. We are fully aware that no such network is likely to exist in any one organism. We use this global network merely as a starting point for successive elimination of reactions, as described in the next section. We generated this global network by merging reactions from two sources, as described in more detail in (Matias Rodrigues and Wagner 2009). The first is the LIGAND reaction database of the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Edwards and Palsson 2000b; Edwards and Palsson 2000a; Goto et al. 2002; Goto et al. 2000; Kanehisa and Goto 2000; Klasson 2004). The second is the complete metabolic reaction set of *E. coli* (iAF1260), which contains 1397 non-transport reactions (Feist et al. 2007). We excluded from this network (i) all reactions involving polymer metabolites of unspecified numbers of monomer units, (ii) general polymerization reactions with uncertain stoichiometry, (ii) reactions involving glycans, due to their complex structure, (iv) reactions with unbalanced stoichiometry, and (v) reactions involving complex metabolites without chemical information about their structure (Matias Rodrigues and Wagner 2009). After these procedures the global network contained 5906 internal (non-transport reactions) and 5030 metabolites.

Essential Reactions and Minimal Networks

A gene whose deletion abolishes the viability of an organism is called an *essential gene*. Analogously, an *essential reaction* in a metabolic network is a reaction that cannot be removed without abolishing the organism's viability in a given

environment. A *minimal* metabolic network is a network from which not a single reaction can be removed without abolishing viability in a given environment. In other words, all reactions of a minimal network are essential in that environment. We emphasize that a minimal network is *not* the smallest possible viable network, which would be difficult to find in a vast space of metabolic networks. We note that there may be multiple minimal networks, which contain different pathways among a set of possible alternate pathway from a nutrient to a biomass precursor. These networks need not have the same size. We also note that the number of essential reactions in any one non-minimal network may be smaller than the size of a minimal network, because non-minimal networks may contain alternate pathways able to by-pass any one reaction.

Generation of Minimal Networks

To analyze the smallest number of reactions that a network needs to have in order to (i) synthesize a given number *B* of biomass molecules on (ii) each of *N* different carbon sources, we created and analyzed many minimal networks with different values of *B* and *N*. Because there are astronomically many combinations of different values of *B* and *N*, and because the FBA approach we use is computationally expensive, we focused on subsets of such combinations, which we created as follows.

First, we created 10 different sets each of *B*=40, *B*=30, and *B*=20 randomly chosen biomass molecules from the set of 63 *E. coli* biomass molecules (excluding the four *'currency'* metabolites mentioned above in the Section 'Biomass Molecules', which are found in all of the sets) (Feist et al. 2007). These sets served as the basis for our analysis of networks that vary in their biosynthetic ability *B*. To arrive at identical

sample sizes for subsequent analyses, we also included 10 "sets" with B=63, that is, each of these sets contained all 63 biomass molecules. In total, we thus created a total of 40 (10×4) sets of biomass molecules.

Second, for each of these 40 sets we determined 4 different sets of nutrients, where each set contained a different number of randomly chosen carbon sources from the list of carbon sources we used (see Section Carbon Sources). Specifically, these sets of nutrients contained N=20, N=40, N=60, and N=80 carbon sources. Thus, up to this second step we had generated a total of 160 (40×4) combinations of sets of biomass molecules and nutrients.

Finally, for each of these 160 combinations, we created 5 minimal metabolic networks. To create each minimal network, we used the following procedure. We started from the global network and sequentially removed individual randomly chosen reactions from it. After each reaction removal we verified that the network was still viable – able to synthesize all B biomass molecules in the set – when each of the N nutrients was provided as the sole carbon source, that is, the network was required to be viable on each carbon source. If that was not the case, we reversed the reaction elimination and removed a different, randomly chosen reaction, until the resulting network was viable. We continued this procedure until no further reactions could be removed from the network without abolishing viability. In this fashion, we generated 800 (160×5) minimal networks. Note that carrying out this procedure repeatedly may not arrive at the same minimal network, because reactions are removed at random. That is, different minimal networks may contain different numbers and different sets of reactions. What unites them is that all their reactions are essential.

Minimal Networks with Isostoichiometric Biomass Composition

We asked whether the stoichiometric composition of biomass, that is, the relative amounts of different biomass molecule in biomass, influences the relationships we explore here. To this end, we generated minimal networks synthesizing given biomass compounds, as described in the previous section, with the difference that these networks synthesized these compounds in equal molarities that is isostoichiometrically. In other words, for the purpose of this analysis we changed the stoichiometric coefficients $\boldsymbol{c}^{T}$ in the biomass growth function of Equation 1 to a value of one. We used the same combinations of sets nutrients and biomass molecules as described above, except that we generated only 2 minimal networks (instead of 5) for each combination, in order to reduce computational cost. In total, we thus analyzed 320 minimal networks with isostoichiometric biomass composition.

Minimal Networks Synthesizing 20 Amino Acids

To study examples of networks that synthesize biochemically related biomass molecules, we studied minimal networks synthesizing only the 20 proteinaceous amino acids. We generated these networks as described in 'Generation of Minimal Networks', except that we did not choose the biomass molecules to be synthesized at random. More specifically, we created 10 sets each of N=20, N=40, N=60, and N=80 randomly chosen carbon sources, and analyzed 8 minimal networks for each set, for a total of 320 networks.

MCMC Sampling and Random Networks

As we mentioned earlier, changing the genotype of a network does not necessarily cause a change in its phenotype. One can take advantage of this property to generate arbitrarily large and unbiased random samples of metabolic networks with any desired property (see Figure 5), such that viable networks with a given number of reactions are sampled uniformly from the space of such networks. Such samples are central to our analysis. To create them we used a procedure built on Markov Chain Monte Carlo (MCMC) sampling described previously (Matias Rodrigues and Wagner 2009; Samal et al. 2010). This procedure fulfills the important detailed balance condition for MCMC sampling. Briefly, the procedure constructs a sequence of metabolic networks, where the next network in the sequence is created from the previous network through a reaction swap. Such a reaction swap consists of the removal of a randomly chosen reaction from the network, followed by the addition of a randomly chosen reaction taken from the global network. We used such reaction swaps, because they preserve the exact number of reactions in a metabolic network, which is important for our analysis. If a reaction swap preserves viability of the network in a given environment, then the swap is accepted, otherwise it is rejected, and a new swap is attempted until a viable genotype is found. We note that subsequent networks in a sequence show autocorrelation in their properties, and are thus not suitable for random sampling. Past work has shown that after 5×10^3 swaps, the autocorrelation of two genotypes becomes negligible (Samal et al. 2010). We therefore started the random walk from *E. coli* metabolic network and after 2.5×10^6 reaction swaps, we sampled networks every 5×10^3 steps in a sequence. Overall, the network samples we

created comprise 1000 networks for each condition we study. We did not subject transport reactions to the reaction swap procedure.

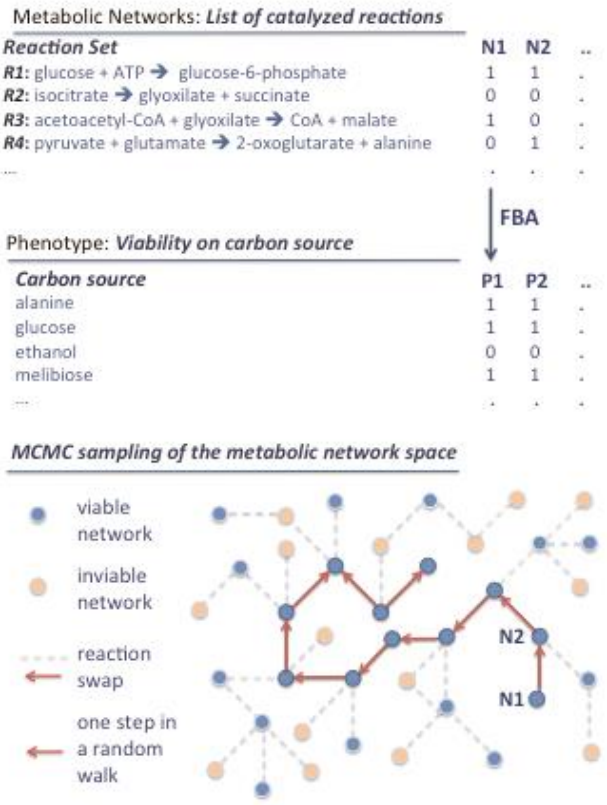

Figure 5. Exploration of a metabolic network space. Metabolic networks can be viewed as subsets of enzyme-catalyzed metabolic reactions in a global reaction set. Formally, they can be represented as binary vectors listing the reactions catalyzed by enzymes in an organism, as indicated for two hypothetical metabolic networks (N1, N2) in the figure. Metabolic phenotypes are computed from metabolic networks using FBA. They can be represented as binary vectors indicating the carbon sources (i.e.: alanine, glucose, melibiose,...) on which a network is viable, that is, on which it can synthesize a given set of (biomass) molecules. Neighboring networks (blue circles linked by edges) differ by a single reaction swap (edges between circles) that leaves the metabolic phenotype unchanged. A reaction swap consists of two changes: one random reaction addition (R4 in the example) and one random reaction deletion (R3 in the example). A series of successful reaction swaps is called a random walk (indicated by red arrows). The Markov Chain Monte Carlo (MCMC) technique allows one to randomly sample networks with a given phenotype by generating long random walks through genotype space, where each step in a walk consists of a reaction swap. The advantage of using reaction swaps is that they leave the number of reactions constant.

To compare our observations from randomly sampled metabolic networks to a reference from biology, it is useful to use a network from a well-studied organism. For this purpose, we used the metabolic network of *E. coli*, because *E. coli* is able to survive in multiple different environments (Touchon et al. 2009; Welch et al. 2002).

Generation of Starting Networks for MCMC Sampling

To initiate the MCMC procedure that generates random samples of networks with a given set of properties, we needed starting networks that have these properties. Specifically, we needed to create networks with a given number of reactions *R*, nutrient flexibility *N*, and biosynthetic ability *B*. To this end, we started with the same 160 sets of *N* nutrients and *B* biomass molecules described in 'Generation of Minimal Networks'. We created for each set 5 viable networks that differ in their number of reactions, that is, they had 400, 500, 600, 700 and 800 reactions. To create these networks, we used the same procedure as for the production of minimal networks, except that we stopped removing reactions when a network with the desired number of reactions had been reached. At the end of this procedure, we had 800 networks (160×5) with various combinations of the 3 network properties. We used each of these networks as starting networks for MCMC sampling, and generated 20 random viable networks from each of the 800 starting networks, for a total of 16000 random networks. This amounts to 200 networks for each different combination of *N*, *B*, and *R*.

We performed all our analyses using MATLAB (7.10.0, The MathWorks Inc., Natick, MA, R2010a) and R (R Development Core Team, 2008).

2.6. Acknowledgements

We thank Olivier Martin, João F. Matias Rodrigues, Vardan Andriasyan and Aditya Barve for helpful discussions. AW acknowledges support through Swiss National Science Foundation grants 315230-129708, as well as through the YeastX project of SystemsX.ch, and the University Priority Research Program in Systems Biology at the University of Zurich.

2.7. References

Antoni, D., Zverlov, V. V, & Schwarz, W. H. (2007). Biofuels from microbes. *Applied Microbiology and Biotechnology*, *77*(1), 23–35. doi:10.1007/s00253-007-1163-x

Bailey, J. E. (1991). Toward a science of metabolic engineering. *Science*, *252*(5013), 1668–1675. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2047876

Becker, S. a, & Palsson, B. Ø. (2005). Genome-scale reconstruction of the metabolic network in Staphylococcus aureus N315: an initial draft to the two-dimensional annotation. *BMC Microbiology*, *5*, 8. doi:10.1186/1471-2180-5-8

Beller, H. R., Goh, E.-B., & Keasling, J. D. (2010). Genes involved in long-chain alkene biosynthesis in Micrococcus luteus. *Applied and Environmental Microbiology*, *76*(4), 1212–23. doi:10.1128/AEM.02312-09

Benner, S. A., & Sismour, A. M. (2005). Synthetic biology. *Nature Reviews Genetics*, *6*(7), 533–543. doi:10.1038/nrg1637

Beste, D. J. V, Hooper, T., Stewart, G., Bonde, B., Avignone-Rossa, C., Bushell, M. E., … McFadden, J. (2007). GSMN-TB: a web-based genome-scale network model of Mycobacterium tuberculosis metabolism. *Genome Biology*, *8*(5), R89. doi:10.1186/gb-2007-8-5-r89

Brooks, S. (1998). Markov chain Monte Carlo method and its application. *Journal of the Royal Statistical Society Series D The Statistician*, *47*(1), 69–100. doi:10.1111/1467-9884.00117

Carr, P. a, & Church, G. M. (2009). Genome engineering. *Nature Biotechnology*, *27*(12), 1151–62. doi:10.1038/nbt.1590

Cases, I., & Lorenzo, V. (2005). Genetically modified organisms for the environment: stories of success and failure and what we have learned from them. *International Microbiology*, *8*(3), 213–222. Retrieved from http://scielo.isciii.es/scielo.php?pid=S1139-67092005000300009&script=sci_abstract&tlng=e

Ciliberti, S., Martin, O. C., & Wagner, A. (2007). Robustness Can Evolve Gradually in Complex Regulatory Gene Networks with Varying Topology. *PLoS Computational Biology*, *3*(2), 10. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17274682

Cooper, G. M. (2008). The Cell. *Cell*, *77*(011600), 67–109. doi:10.1007/978-0-387-79240-8

Covert, M. W., Schilling, C. H., & Palsson, B. (2001). Regulation of gene expression in flux balance models of metabolism. *Journal of Theoretical Biology*, *213*(1), 73–88. doi:10.1006/jtbi.2001.2405

Diaconis, P. (2008). The Markov chain Monte Carlo revolution. *Bulletin of the American Mathematical Society*, *46*(2), 179–205. doi:10.1090/S0273-0979-08-01238-X

Domingues, A., Vinga, S., & Lemos, J. M. (2010). Optimization strategies for metabolic networks. *BMC Systems Biology*, *4*(1), 113. Retrieved from http://www.biomedcentral.com/1752-0509/4/113

Dueñas-Sánchez, R., Codón, A. C., Rincón, A. M., & Benítez, T. (2010). Increased biomass production of industrial bakers' yeasts by overexpression of Hap4 gene. *International Journal of Food Microbiology*, *143*(3), 150–160. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20832886

Ebenhöh, O., & Heinrich, R. (2001). Evolutionary optimization of metabolic pathways. Theoretical reconstruction of the stoichiometry of ATP and NADH producing systems. *Bulletin of Mathematical Biology*, *63*(1), 21–55. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11146883

Edwards, J. S., & Palsson, B. O. (2000a). Metabolic flux balance analysis and the in silico analysis of Escherichia coli K-12 gene deletions. *BMC Bioinformatics*, *1*, 1. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=29061&tool=pmcentrez&rendertype=abstract

Edwards, J. S., & Palsson, B. O. (2000b). Robustness analysis of the Escherichia coli metabolic network. *Biotechnology Progress*, *16*(6), 927–39. doi:10.1021/bp0000712

Expósito, O., Bonfill, M., Moyano, E., Onrubia, M., Mirjalili, M. H., Cusidó, R. M., & Palazón, J. (2009). Biotechnological production of taxol and related taxoids: current state and prospects. *Anticancer Agents in Medicinal Chemistry*, *9*(1), 109–121. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19149486

Feist, A. M., Henry, C. S., Reed, J. L., Krummenacker, M., Joyce, A. R., Karp, P. D., … Palsson, B. Ø. (2007). A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Molecular Systems Biology*, *3*(121), 121. doi:10.1038/msb4100155

Feist, A. M., & Palsson, B. O. (2010). The biomass objective function. *Current Opinion in Microbiology*, *13*(3), 344–349. doi:10.1016/j.mib.2010.03.003

Feist, A. M., Scholten, J. C. M., Palsson, B. Ø., Brockman, F. J., & Ideker, T. (2006). Modeling methanogenesis with a genome-scale metabolic reconstruction of Methanosarcina barkeri. *Molecular Systems Biology*, *2*, 2006.0004. doi:10.1038/msb4100046

Fong, S., & Joyce, A. (2005). Parallel adaptive evolution cultures of Escherichia coli lead to convergent growth phenotypes with different gene expression states. *Genome Research*, *15*(858), 1365–1372. doi:10.1101/gr.3832305.15

Fong, S. S., Marciniak, J. Y., & Palsson, B. Ø. (2003). Description and Interpretation of Adaptive Evolution of Escherichia coli K-12 MG1655 by Using a Genome-Scale In Silico Metabolic Model. *Journal Of Bacteriology*, *185*(21), 6400–6408. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=219384&tool=pmcentrez&rendertype=abstract

Forster, A. C., & Church, G. M. (2006). Towards synthesis of a minimal cell. *Molecular Systems Biology*, *2*, 45. doi:10.1038/msb4100090

Garcia Sanchez, R., Karhumaa, K., Fonseca, C., Sànchez Nogué, V., Almeida, J. R., Larsson, C. U., … Gorwa-Grauslund, M. F. (2010). Improved xylose and arabinose utilization by an industrial recombinant Saccharomyces cerevisiae strain using evolutionary engineering. *Biotechnology for Biofuels*, *3*(1), 13.

Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2908073&tool=pmcentrez&rendertype=abstract

Gibson, D. G., Benders, G. a, Andrews-Pfannkoch, C., Denisova, E. a, Baden-Tillson, H., Zaveri, J., … Smith, H. O. (2008). Complete chemical synthesis, assembly, and cloning of a Mycoplasma genitalium genome. *Science (New York, N.Y.)*, *319*(5867), 1215–20. doi:10.1126/science.1151721

Gibson, D. G., Glass, J. I., Lartigue, C., Noskov, V. N., Chuang, R.-Y., Algire, M. a, … Venter, J. C. (2010). Creation of a bacterial cell controlled by a chemically synthesized genome. *Science (New York, N.Y.)*, *329*(5987), 52–6. doi:10.1126/science.1190719

Gil, R., Silva, F. J., Zientz, E., Delmotte, F., González-Candelas, F., Latorre, A., … Moya, A. (2003). The genome sequence of Blochmannia floridanus: comparative analysis of reduced genomes. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(16), 9388–9393. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=170928&tool=pmcentrez&rendertype=abstract

Glass, J. I., Assad-Garcia, N., Alperovich, N., Yooseph, S., Lewis, M. R., Maruf, M., … Venter, J. C. (2006). Essential genes of a minimal bacterium. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(2), 425–30. doi:10.1073/pnas.0510013103

Gonzalez, O. R., Küper, C., Jung, K., Naval, P. C., & Mendoza, E. (2007). Parameter estimation using Simulated Annealing for S-system models of biochemical networks. *Bioinformatics*, *23*(4), 480–486. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17038344

Goto, S., Nishioka, T., & Kanehisa, M. (2000). LIGAND: chemical database of enzyme reactions. *Nucleic Acids Research*, *28*(1), 380–2. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=102410&tool=pmcentrez&rendertype=abstract

Goto, S., Okuno, Y., Hattori, M., Nishioka, T., & Kanehisa, M. (2002). LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Research*, *30*(1), 402–4. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=99090&tool=pmcentrez&rendertype=abstract

Handorf, T., Christian, N., Ebenhöh, O., & Kahn, D. (2008). An environmental perspective on metabolism. *Journal of Theoretical Biology*, *252*(3), 530–537. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18086477

Hillenmeyer, M. E., Fung, E., Wildenhain, J., Pierce, S. E., Hoon, S., Lee, W., … Giaever, G. (2008). The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science (New York, N.Y.)*, *320*(5874), 362–5. doi:10.1126/science.1150021

Ho, N. W. Y., Chen, Z., & Brainard, A. P. (1998). Genetically engineered Saccharomyces yeast capable of effective cofermentation of glucose and xylose. *Applied and Environmental Microbiology*, *64*(5), 1852–1859. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=106241&tool=pmcentrez&rendertype=abstract

Julius, A. A., Imielinski, M., & Pappas, G. J. (2008). Metabolic networks analysis using convex optimization. *2008 47th IEEE Conference on Decision and Control*, 762–767. doi:10.1109/CDC.2008.4739111

Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, *28*(1), 27–30. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=102409&tool=pmcentrez&rendertype=abstract

Kauffman, K. J., Prakash, P., & Edwards, J. S. (2003). Advances in flux balance analysis. *Current Opinion in Biotechnology*, *14*(5), 491–496. doi:10.1016/j.copbio.2003.08.001

Keasling, J. D. (2010). Manufacturing Molecules Through Metabolic Engineering. *Science*, *330*(6009), 1355–1358. doi:10.1126/science.1193990

Kim, J., & Reed, J. L. (2010). OptORF: Optimal metabolic and regulatory perturbations for metabolic engineering of microbial strains. *BMC Systems Biology*, *4*(1), 53. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2887412&tool=pmcentrez&rendertype=abstract

Klasson, L. (2004). Evolution of minimal-gene-sets in host-dependent bacteria. *Trends in Microbiology*, *12*(1), 37–43. doi:10.1016/j.tim.2003.11.006

Kuwahara, H., Yoshida, T., Takaki, Y., Shimamura, S., Nishi, S., Harada, M., … Maruyama, T. (2007). Reduced genome of the thioautotrophic intracellular symbiont in a deep-sea clam, Calyptogena okutanii. *Current Biology : CB*, *17*(10), 881–6. doi:10.1016/j.cub.2007.04.039

Lee, S. K., Chou, H., Ham, T. S., Lee, T. S., & Keasling, J. D. (2008). Metabolic engineering of microorganisms for biofuels production: from bugs to synthetic biology to fuels. *Current Opinion in Biotechnology*, *19*(6), 556–63. doi:10.1016/j.copbio.2008.10.014

Leuchtenberger, W., Huthmacher, K., & Drauz, K. (2005). Biotechnological production of amino acids and derivatives: current status and prospects. *Applied Microbiology and Biotechnology*, *69*(1), 1–8. doi:10.1007/s00253-005-0155-y

Li, X. R. (1992). Generation of random points uniformly distributed in hyperellipsoids. In *Proc First IEEE Conference on Control Applications* (pp. 654–658 vol.2). doi:10.1109/CCA.1992.269770

Liang, J. C., Bloom, R. J., & Smolke, C. D. (2011). Engineering Biological Systems with Synthetic RNA Molecules. *Molecular Cell*, *43*(6), 915–926. doi:10.1016/j.molcel.2011.08.023

Liu, E., & Hu, Y. (2010). Construction of a xylose-fermenting Saccharomyces cerevisiae strain by combined approaches of genetic engineering, chemical mutagenesis and evolutionary adaptation. *Biochemical Engineering Journal*, *48*(2), 204–210. doi:10.1016/j.bej.2009.10.011

Madhavan, A., Tamalampudi, S., Srivastava, A., Fukuda, H., Bisaria, V. S., & Kondo, A. (2009). Alcoholic fermentation of xylose and mixed sugars using recombinant Saccharomyces cerevisiae engineered for xylose utilization. *Applied Microbiology and Biotechnology*, *82*(6), 1037–47. doi:10.1007/s00253-008-1818-2

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, *18*(1), 50–60. doi:10.1214/aoms/1177730491

Matias Rodrigues, J. F., & Wagner, A. (2009). Evolutionary plasticity and innovations in complex metabolic reaction networks. *PLoS Computational Biology*, *5*(12), e1000613. doi:10.1371/journal.pcbi.1000613

Metzger, P., & Largeau, C. (2005). Botryococcus braunii: a rich source for hydrocarbons and related ether lipids. *Applied Microbiology and Biotechnology*, *66*(5), 486–496. doi:10.1007/s00253-004-1779-z

Mira, A., Ochman, H., & Moran, N. A. (2001). Deletional bias and the evolution of bacterial genomes. *Trends in Genetics*, *17*(10), 589–596. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11585665

Mizoguchi, H., Sawano, Y., Kato, J., & Mori, H. (2008). Superpositioning of deletions promotes growth of Escherichia coli with a reduced genome. *DNA Research : An International Journal for Rapid Publication of Reports on Genes and Genomes*, *15*(5), 277–84. doi:10.1093/dnares/dsn019

Morot-Gaudry, J. F., Job, D., & Lea, P. J. (2001). Amino acid metabolism. In *Medical Biochemistry* (Vol. 28, pp. 299–329). Springer Verlag. Retrieved from http://eprints.lancs.ac.uk/10808/

Mukhopadhyay, A., Redding, A. M., Rutherford, B. J., & Keasling, J. D. (2008). Importance of systems biology in engineering microbes for biofuel production. *Current Opinion in Biotechnology*, *19*(3), 228–34. doi:10.1016/j.copbio.2008.05.003

Murtas, G. (2007). Question 7: construction of a semi-synthetic minimal cell: a model for early living cells. *Origins of Life and Evolution of the Biosphere : The Journal of the International Society for the Study of the Origin of Life*, *37*(4-5), 419–22. doi:10.1007/s11084-007-9090-5

Nakabachi, A., Yamashita, A., Toh, H., Ishikawa, H., Dunbar, H. E., Moran, N. A., & Hattori, M. (2008). Bacterial Endosymbiont Carsonella. *October*, (March 2007), 3–5.

Neal, R. M. (1993). Probabilistic Inference Using Markov Chain Monte Carlo Methods. *Intelligence*, *62*(September), 144. doi:10.1.1.46.8183

Nevoigt, E. (2008). Progress in Metabolic Engineering of Saccharomyces cerevisiae. *Microbiology and Molecular Biology Reviews MMBR*, *72*(3), 379–412. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2546860&tool=pmcentrez&rendertype=abstract

Nishikawa, T., Gulbahce, N., & Motter, A. E. (2008). Spontaneous reaction silencing in metabolic optimization. *PLoS Computational Biology*, *4*(12), e1000236. doi:10.1371/journal.pcbi.1000236

Oberhardt, M. A., Palsson, B. Ø., & Papin, J. A. (2009). Applications of genome-scale metabolic reconstructions. *Molecular Systems Biology*, *5*(320), 320. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19888215

Oh, Y.-K., Palsson, B. O., Park, S. M., Schilling, C. H., & Mahadevan, R. (2007). Genome-scale reconstruction of metabolic network in Bacillus subtilis based on high-throughput phenotyping and gene essentiality data. *The Journal of Biological Chemistry*, *282*(39), 28791–9. doi:10.1074/jbc.M703759200

Orth, J. D., Thiele, I., & Palsson, B. Ø. (2010). What is flux balance analysis? *Nature Biotechnology*, *28*(3), 245–8. doi:10.1038/nbt.1614

Patil, K. R., Rocha, I., Förster, J., & Nielsen, J. (2005). Evolutionary programming as a platform for in silico metabolic engineering. *BMC Bioinformatics*, *6*(1), 308. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16375763

Paul G. Higgs, T. K. A. (2005). *Bioinformatics and molecular evolution.* (p. 384). Oxford, UK: John Wiley & Sons, Ltd. doi:10.1002/cfg.486

Price, N. D., Reed, J. L., & Palsson, B. Ø. (2004). Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nature Reviews. Microbiology*, *2*(11), 886–97. doi:10.1038/nrmicro1023

Purnick, P. E. M., & Weiss, R. (2009). The second wave of synthetic biology: from modules to systems. *Nature Reviews. Molecular Cell Biology*, *10*(6), 410–22. doi:10.1038/nrm2698

Raab, A. M., Hlavacek, V., Bolotina, N., & Lang, C. (2011). Shifting the Fermentative/Oxidative Balance in Saccharomyces cerevisiae by Transcriptional Deregulation of Snf1 via Overexpression of the Upstream Activating Kinase Sak1p. *Applied and Environmental Microbiology*, *77*(6), 1981–1989. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21257817

Raghevendran, V., Patil, K. R., Olsson, L., & Nielsen, J. (2006). Hap4 is not essential for activation of respiration at low specific growth rates in Saccharomyces cerevisiae. *The Journal of Biological Chemistry*, *281*(18), 12308–12314. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16522629

Rasmussen, S., Chen, L., Deamer, D., Krakauer, D. C., Packard, N. H., Stadler, P. F., & Bedau, M. a. (2004). Evolution. Transitions from nonliving to living matter. *Science (New York, N.Y.)*, *303*(5660), 963–5. doi:10.1126/science.1093669

Riley, M. (1997). Genes and proteins of Escherichia coli K-12 (GenProtEC). *Nucleic Acids Research*, *25*(1), 51–52. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=146413&tool=pmcentrez&rendertype=abstract

Robert, C. P., & Casella, G. (2004). *Monte Carlo Statistical Methods*. *Book* (Vol. 96, p. 645). Springer. Retrieved from http://worldcat.org/isbn/0-387-21239-6

Rocha, I., Maia, P., Evangelista, P., Vilaça, P., Soares, S., Pinto, J. P., … Rocha, M. (2010). OptFlux: an open-source software platform for in silico metabolic engineering. *BMC Systems Biology*, *4*(1), 45. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2864236&tool=pmcentrez&rendertype=abstract

Rude, M. a, & Schirmer, A. (2009). New microbial fuels: a biotech perspective. *Current Opinion in Microbiology*, *12*(3), 274–81. doi:10.1016/j.mib.2009.04.004

Samal, A., Matias Rodrigues, J. F., Jost, J., Martin, O. C., & Wagner, A. (2010). Genotype networks in metabolic reaction spaces. *BMC Systems Biology*, *4*, 30. doi:10.1186/1752-0509-4-30

Savage, D. F., Way, J., & Silver, P. A. (2008). Defossiling fuel: how synthetic biology can transform biofuel production. *ACS Chemical Biology*, *3*(1), 13–16. Retrieved from http://pubs.acs.org/doi/abs/10.1021/cb700259j

Schellenberger, J., Que, R., Fleming, R. M. T., Thiele, I., Orth, J. D., Feist, A. M., … Palsson, B. Ø. (2007). Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nature Protocols*, *2*(3), 1290–1307. doi:10.1038/nprot.2011.308

Schilling, C. H., Covert, M. W., Famili, I., Church, G. M., Edwards, J. S., & Palsson, B. O. (2002). Genome-scale metabolic model of Helicobacter pylori 26695. *Journal Of Bacteriology*, *184*(16), 4582–4593. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=12142428

Schirmer, a., Rude, M. a., Li, X., Popova, E., & del Cardayre, S. B. (2010). Microbial Biosynthesis of Alkanes. *Science*, *329*(5991), 559–562. doi:10.1126/science.1187936

Schmidt, C. W. (2010). Synthetic Biology: Environmental Health Implications of a New Field. *Environmental Health Perspectives*, *118*(3), A118–A123. Retrieved from http://apps.isiknowledge.com/full_record.do?product=WOS&search_mode=GeneralSearch&qid=1&SID=V2A83giKpi2Hk1b5L1L&page=1&doc=35

Segrè, D., Vitkup, D., & Church, G. M. (2002). Analysis of optimality in natural and perturbed metabolic networks. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(23), 15112–7. doi:10.1073/pnas.232349399

Shendure, J., Mitra, R. D., Varma, C., & Church, G. M. (2004). Advanced sequencing technologies: methods and goals. *Nature Reviews. Genetics*, *5*(5), 335–44. doi:10.1038/nrg1325

Smallbone, K., & Simeonidis, E. (2009). Flux balance analysis: a geometric perspective. *Journal of Theoretical Biology*, *258*(2), 311–5. doi:10.1016/j.jtbi.2009.01.027

Smith, H. O., Hutchison, C. a, Pfannkoch, C., & Venter, J. C. (2003). Generating a synthetic genome by whole genome assembly: phiX174 bacteriophage from synthetic oligonucleotides. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(26), 15440–5. doi:10.1073/pnas.2237126100

Smolke, C. D., & Silver, P. a. (2011). Informing biological design by integration of systems and synthetic biology. *Cell*, *144*(6), 855–9. doi:10.1016/j.cell.2011.02.020

Sprinzak, D., & Elowitz, M. B. (2005). Reconstruction of genetic circuits. *Nature*, *438*(7067), 443–8. doi:10.1038/nature04335

Steen, E. J., Chan, R., Prasad, N., Myers, S., Petzold, C. J., Redding, A., … Keasling, J. D. (2008). Metabolic engineering of Saccharomyces cerevisiae for the production of n-butanol. *Microbial Cell Factories*, *7*, 36. doi:10.1186/1475-2859-7-36

Steen, E. J., Kang, Y., Bokinsky, G., Hu, Z., Schirmer, A., McClure, A., … Keasling, J. D. (2010). Microbial production of fatty-acid-derived fuels and chemicals from plant biomass. *Nature*, *463*(7280), 559–62. doi:10.1038/nature08721

Stephanopoulos, G., & Vallino, J. J. (1991). Network rigidity and metabolic engineering in metabolite overproduction. *Science*, *252*(5013), 1675–1681. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1904627

Thomas, G. H., Zucker, J., Macdonald, S. J., Sorokin, A., Goryanin, I., & Douglas, A. E. (2009). A fragile metabolic network adapted for cooperation in the symbiotic bacterium Buchnera aphidicola. *BMC Systems Biology*, *3*, 24. doi:10.1186/1752-0509-3-24

Tomshine, J., & Kaznessis, Y. N. (2006). Optimization of a Stochastically Simulated Gene Network Model via Simulated Annealing. *Biophysical Journal*, *91*(9), 3196–3205. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16920827

Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., … Denamur, E. (2009). Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. *PLoS Genetics*, *5*(1), e1000344. doi:10.1371/journal.pgen.1000344

Vandamme, E. J. (1992). Production of vitamins, coenzymes and related biochemicals by biotechnological processes. *Journal of Chemical Technology and Biotechnology Oxford Oxfordshire 1986*, *53*(4), 313–327. Retrieved from http://doi.wiley.com/10.1002/chin.199229310

Wang, Z., & Zhang, J. (2009). Abundant Indispensable Redundancies in Cellular Metabolic Networks. *Genome Biology and Evolution*, *2009*(0), 23–33. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2817398&tool=pmcentrez&rendertype=abstract

Weber, J., Kayser, A., & Rinas, U. (2005). Metabolic flux analysis of Escherichia coli in glucose-limited continuous culture. II. Dynamic response to famine and feast, activation of the methylglyoxal pathway and oscillatory behaviour. *Microbiology (Reading, England)*, *151*(Pt 3), 707–16. doi:10.1099/mic.0.27482-0

Welch, R. a, Burland, V., Plunkett, G., Redford, P., Roesch, P., Rasko, D., … Blattner, F. R. (2002). Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(26), 17020–4. doi:10.1073/pnas.252529799

Wessely, F., Bartl, M., Guthke, R., Li, P., Schuster, S., & Kaleta, C. (2011). Optimal regulatory strategies for metabolic pathways in Escherichia coli depending on protein costs. *Molecular Systems Biology*, *7*(515), 515. Retrieved from http://www.nature.com/doifinder/10.1038/msb.2011.46

Wisselink, H. W., Toirkens, M. J., Wu, Q., Pronk, J. T., & van Maris, A. J. a. (2009). Novel evolutionary engineering approach for accelerated utilization of glucose, xylose, and arabinose mixtures by engineered Saccharomyces cerevisiae strains. *Applied and Environmental Microbiology*, *75*(4), 907–14. doi:10.1128/AEM.02268-08

Yang, L., Mahadevan, R., & Cluett, W. (2008). A bilevel optimization algorithm to identify enzymatic capacity constraints in metabolic networks. *Computers & Chemical Engineering*, *32*(9), 2072–2085. doi:10.1016/j.compchemeng.2007.10.015

Yarmush, M. L., & Banta, S. (2003). Metabolic engineering: advances in modeling and intervention in health and disease. *Annual Review of Biomedical Engineering*, *5*, 349–81. doi:10.1146/annurev.bioeng.5.031003.163247

2.8. Supplementary Material

Supplementary Tables

L-alanine	
L-arginine	potassium
L-asparagine	ammonium
L-aspartate	magnesium
L-cysteine	calcium

L-glutamine	reduced iron
L-glutamate	iron trication
glycine	copper
L-histidine	manganese
L-isoleucine	molybdenum
L-leucine	cobalt
L-lysine	zinc
L-methionine	chloride
L-phenylalanine	sulfate
L-proline	water
L-serine	coenzyme-A
L-threonine	NAD
L-tryptophan	NADP
L-tyrosine	FAD
L-valine	5,6,7,8-tetrahydrofolate
datp	5,10-methylenetetrahydrofolate
dttp	10-formyltetrahydrofolate
dgtp	thiamine diphosphate
dctp	pyridoxal 5'-phosphate
CTP	protoheme
GTP	siroheme
UTP	undecaprenyl diphosphate
ATP	S-sdenosyl-L-methionine
murein disaccharide	2-octaprenyl-6-hydroxyphenol
KDO(2)-lipid IV(A)	riboflavin
Phosphatidylethanolamine (dihexadecanoyl, n-C16:0)	phosphatidylethanolamine (dihexadecanoyl, n-C16:1)
Phosphatidylethanolamine phosphate (dihexadecanoyl, n-C16:0)	phosphatidylethanolamine phosphate (dihexadecanoyl, n-C16:1)

Table S1. List of *E. coli's* Biomass Compounds [57]

1,4-alpha-D-glucan	UMP	Lactose
3-hydroxycinnamic	Glycerol	Butyrate
3-(3-hydroxy-phenyl)propionate	alpha-D-Ribose	Hexadecanoate
Phosphate	D-Ribose	D-Lactate
AMP	Fumarate	D-Gluconate
Pyruvate	D-Galactose	L-Arabinose
L-Glutamate	IMP	Hypoxanthine
2-Oxoglutarate	D-Alanine	N-Acetylneuraminate
UDPglucose	Putrescine	D-Mannose
D-Glucose	N-Acetyl-D-glucosamine	Inosine
Acetate	GMP	Uridine
Glycine	Adenine	L-Xylulose
L-Alanine	L-Proline	D-Glucosamine
Succinate	L-Malate	Deoxyguanosine
UDP-N-acetyl-D-glucosamine	L-Asparagine	D-Galacturonate
L-Aspartate	Citrate	4-Aminobutanoate
Reduced	D-Mannose	D-Glucosamine

UDPgalactose	Glycolate	dAMP
CMP	Propionate	dGMP
Formate	Acetoacetate	dTMP
Sulfate	UDP-D-glucuronate	dUMP
L-Arginine	Agmatine	Xanthine
L-Glutamine	D-Xylose	Guanosine
L-Serine	Dihydroxyacetone	D-Mannitol
Formaldehyde	L-Lactate	alpha-D-Galactose
L-Ascorbate	L-Threonine	Ethanol
L-Tryptophan	Ethanolamine	Cytidine
Acetaldehyde	D-Glucuronate	Propanal
D-Fructose	UDP-N-acetyl-D-galactosamine	D-Malate
Sucrose	2-Dehydro-3-deoxy-D-gluconate	L-Rhamnose
D-Glucose	Maltose	Deoxyuridine
Glycerol	Adenosine	Deoxyadenosine
D-Fructose	Thymidine	D-Glyceraldehyde
L-Cysteine	dCMP	N-Acetyl-D-mannosamine
D-Glucose	Guanine	Xanthosine
sn-Glycero-3-phosphocholine	2',3'-Cyclic	octadecanoate
D-Serine	2',3'-Cyclic	Allantoin
L-Idonate	Dodecanoate	Decanoate
D-Cysteine	N-Acetylmuramate	Hexanoate
D-Sorbitol	Glycerol	Ornithine
D-Glucarate	Glycerophosphoglycerol	Galactitol
D-Galactarate	N-Acetyl-D-glucosamine	Xanthosine
D-Galactonate	D-Glucuronate	Maltotriose
Deoxycytidine	Melibiose	Maltohexaose
L-tartrate	Deoxyinosine	Maltotetraose
D-Fructuronate	Phenylpropanoate	2',3'-Cyclic
D-Alanyl-D-alanine	3'-cmp	butanesulfonate
O-Phospho-L-serine	3'-GMP	ethanesulfonate
L-Fucose	2',3'-Cyclic	fructoselysine
5-Dehydro-D-gluconate	dIMP	Glycerophosphoserine
Trehalose	Fe(III)dicitrate	Galactonate
sn-Glycero-3-phospho-1-inositol	2,3-diaminopropionate	Hexadecenoate
sn-Glycero-3-phosphoethanolamine	octanoate	Maltopentaose
Ammonium	tetradecanoate	octadecenoate
3'-AMP	2(alpha-D-Mannosyl)-D-glycerate	L-Prolinylglycine
3'-UMP	L-Threonine	psicoselysine
Cys-Gly	L-Lyxose	tetradecenoate
L-alanine-D-glutamate-meso-2,6-diaminoheptanedioate	N-Acetyl-D-glucosamine(anhydrous)N-Acetylmuramic	L-alanine-D-glutamate-meso-2,6-diaminoheptanedioate-D-alanine
D-Allose		

Table S2. List of carbon sources used in this study

Examples of reactions required to synthesize additional biomass molecules

Biomass Molecule	Required Additional Reactions
2-octaprenyl-6-hydroxyphenol	2-octaprenylphenol hydroxylase, octaprenyl-hydroxybenzoate decarboxylase, hydroxybenzoate octaprenyltransferase, chorismate pyruvate lyase, octaprenyl pyrophosphate synthase
arginine	N-acetylglutamate synthase, N-acetyl-g-glutamyl-phosphate reductase, acetylglutamate kinase, acetylornithine transaminase, acetylornithine deacetylase, argininosuccinate lyase, argininosuccinate synthase, ornithine carbamoyltransferase
asparagine	asparagine synthase
coenzyme A	2-dehydropantoate 2-reductase, 3-methyl-2-oxobutanoate hydroxymethyltransferase, phosphopantothenate-cysteine ligase, phosphopantothenoylcysteine decarboxylase, pantothenate synthase, aspartate 1-decarboxylase, dephospho-CoA kinase, pantetheine-phosphate adenylyltransferase, pantothenate kinase
dATP	ribonucleoside-triphosphate reductase (ATP)
dCTP	ribonucleoside-triphosphate reductase (CTP)
dGTP	nucleoside-diphosphate kinase (ATP:dGDP), ribonucleoside-diphosphate reductase (GDP)
dTTP	nucleoside-diphosphate kinase (ATP:dTDP), dTMP kinase, uridylate kinase (dUMP), thymidylate synthase, ribonucleoside-diphosphate reductase (UDP)
FAD	FMN adenylyltransferase, riboflavin kinase
histidine	ATP phosphoribosyltransferase, imidazoleglycerol-phosphate dehydratase, imidazole-glycerol-3-phosphate synthase, histidinol-phosphate transaminase, histidinol-phosphatase, histidinol dehydrogenase, phosphoribosyl-AMP cyclohydrolase, phosphoribosyl-ATP pyrophosphatase, 1-imidazole-4-carboxamide isomerase
isoleucine	dihydroxy-acid dehydratase, ketol-acid reductoisomerase, 2-aceto-2-hydroxybutanoate synthase, isoleucine transaminase, L-threonine deaminase
KDO(2)-lipid IV(A)	UDP-3-O-glucosamine acyltransferase, UDP-N-acetylglucosamine acyltransferase, 3-deoxy -D-manno-octulosonic -acid 8-phosphate, tetraacyldisaccharide 4'kinase, 3-deoxy-D-manno-octulosonic acid transferase, 3-deoxy-manno-octulosonate cytidylyltransferase, 3-deoxy-manno-octulosonate-8-phosphatase, UDP-sugar hydrolase, UDP-3-O-acetylglucosamine deacetylase, lipid A disaccaride synthase, arabinose-5-phosphate isomerase
leucine	2-isopropylmalate hydratase, 3-isopropylmalate dehydrogenase, 3-isopropylmalate dehydratase, 2-Oxo-4-methyl-3-carboxypentanoate decarboxylation, 2-isopropylmalate synthase, leucine transaminase
lysine	diaminopimelate decarboxylase
acetyl-CoA	2-octaprenylphenol hydroxylase, octaprenyl-hydroxybenzoate decarboxylase, hydroxybenzoate octaprenyltransferase, chorismate pyruvate lyase, octaprenyl pyrophosphate synthase
alanine	phosphogluconate dehydrogenase
aspartate	malic enzyme
chorismate	2-dehydropantoate 2-reductase, 3-methyl-2-oxobutanoate hydroxymethyltransferase, phosphopantothenate-cysteine ligase, phosphopantothenoylcysteine decarboxylase, pantothenate synthase, aspartate 1-decarboxylase, dephospho-CoA kinase, pantetheine-phosphate adenylyltransferase, pantothenate kinase
protoheme	Valine-pyruvate aminotransferase
putrescine	Undecaprenyl diphosphate synthase

Table S3. Examples of reactions required to synthesize additional biomass molecules. The table contains 20

arbitrary biomass molecules (left), and a list of reactions that are required to synthesize the molecule in a random minimal network (in addition to the reactions that the network needed to synthesize other biomass molecules). The analysis is based on minimal networks that are required (i) to synthesize 62 *E. coli* biomass molecules and (ii) to be viable on glucose. The Table illustrates that the number of additional reactions needed depends on the biomass molecule. (It may also depend on other reactions in a network, but for each biomass molecule results from only one network are shown.)

Examples of reactions required to utilize additional carbon sources

Carbon Source	Required Additional Reactions
D-galactose	galactokinase, UDPglucose--hexose-1-phosphate uridylyltransferase, UDPglucose 4-epimerase
glycerophosphoserine	glycerophosphodiester phosphodiesterase
1,4-alpha-D-glucan	maltodextrin glucosidase
2(alpha-D-Mannosyl)-D-glycerate	2(alpha-D-Mannosyl-6-phosphate)-D-glycerate hydrolase
L-ascorbate	3-keto-L-gulonate 6-phosphate decarboxylase, L-ribulose-phosphate 4-epimerase, L-xylulose 5-phosphate 3-epimerase
agmatine	agmatinase
IMP	5'-nucleotidase
dAMP	deoxyadenosine deaminase, purine-nucleoside phosphorylase
L-tartrate	L(+)-tartrate dehydratase
deoxyguanosine	purine-nucleoside phosphorylase (Deoxyguanosine)
2',3'-Cyclic	2',3'-Cyclic UMP phasphatase
phenylpropanoate	4-hydroxy-2-oxopentanoate aldolase, 2,3-dihydroxyphenylpropionate dehydrogenase, diaminohydroxyphosphoribosylaminopyrimidine deaminase, 2,3-dihydroxypheylpropionate 1,2-dioxygenase, phenylpropanoate Dioxygenase, 2,3-dihydroxypheylpropionate 1,2-dioxygenase, 2-hydroxy-6-ketonona-2,4-dienedioic acid hydrolase, 2-oxopent-4-enoate hydratase
glycerophosphoglycerol	glycerophosphodiester phosphodiesterase
L-idonate	L-idonate 5-dehydrogenase
D-ribose	ribokinase
lactose	b-galactosidase
D-galactarate	5-dehydro-4-deoxyglucarate aldolase, galactarate dehydratase
hypoxanthine	ureidoglycolate hydrolase, malate synthase, allantoinase, 5'-nucleotidase (GMP), purine-nucleoside phosphorylase (Guanosine), guanine deaminase, purine-nucleoside phosphorylase (Inosine), L-serine deaminase
N-acetyl-D-glucosamine 1-phosphate	glycerophosphodiester phosphodiesterase
asparagine	asparaginase

Table S4. Examples of reactions required to utilize additional carbon sources. The table contains arbitrary carbon sources (left) and a list of reactions that are required to utilize the carbon source in a random minimal network (in addition to the reactions that the network needs to utilize other carbon sources). The analysis is based on minimal networks that were required (i) to synthesize all *E. coli* biomass molecules and (ii) to be viable on 30 other carbon sources. The Table illustrates that the number of additional reactions needed depends on the carbon source. (It may also depend on other reactions in a network, but for each carbon source results for only one network are shown.)

Supplementary Figures

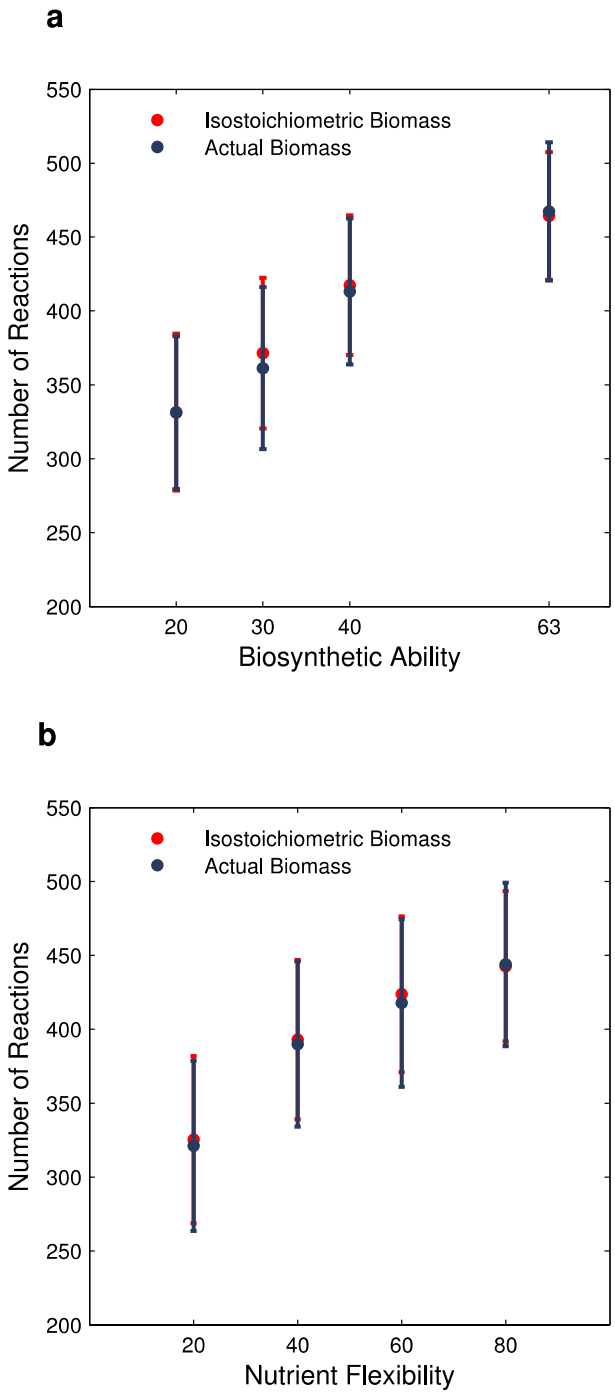

Figure S1. Biomass stoichiometry does not affect the number of reactions in a minimal network. The vertical axis shows the number of reactions in minimal networks as a function of a) biosynthetic ability and b) nutrient flexibility. Dots and lengths of error bars correspond to means and one standard deviation. Each blue dot of size 200 indicates networks with the biomass stoichiomery of *E. coli* [57]. Each red dot of size 80 indicates networks with isostoichiometric biomass (see Methods). Red and blue means do not differ from each other significantly (P>0.40, Mann-Whitney U-test).

Biosynthetic Flux Distribution

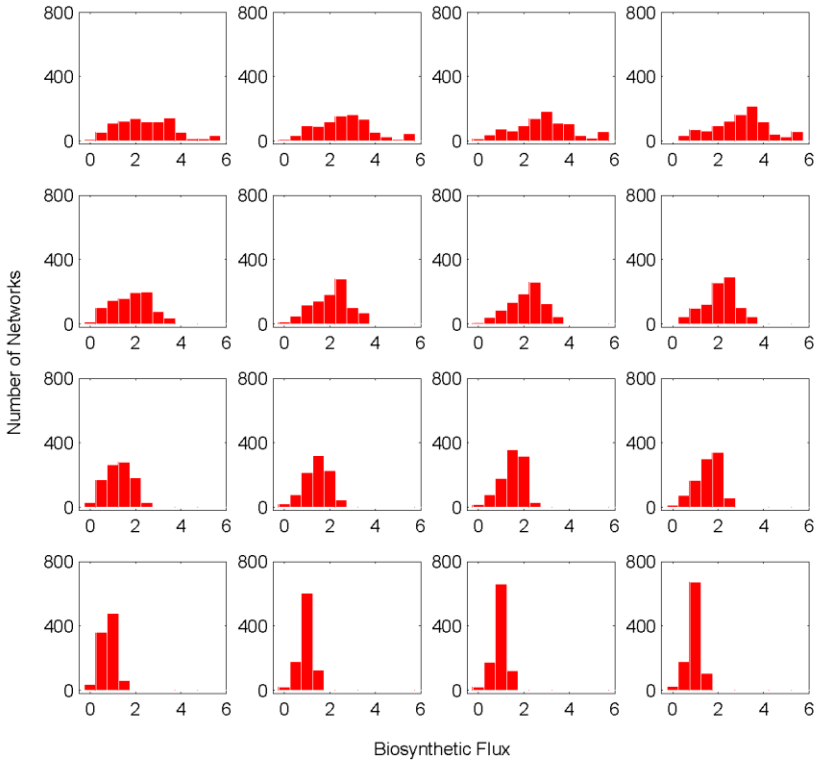

Figure S2. Biosynthetic flux distribution. The flux distributions in units of mmoles per g DW per hour are shown for each combination of biosynthetic ability (*B*=20 first row, *B*=30 second row, *B*=40 third row, *B*=63 last row) and nutrient flexibility (*N*=20 first column, *N*=30 second column, *N*=40 third column, *N*=63 last column) that we examined. Data are based on 16,000 random viable network, as described in methods (1000 networks per panel).

Distribution of number *R* of reactions

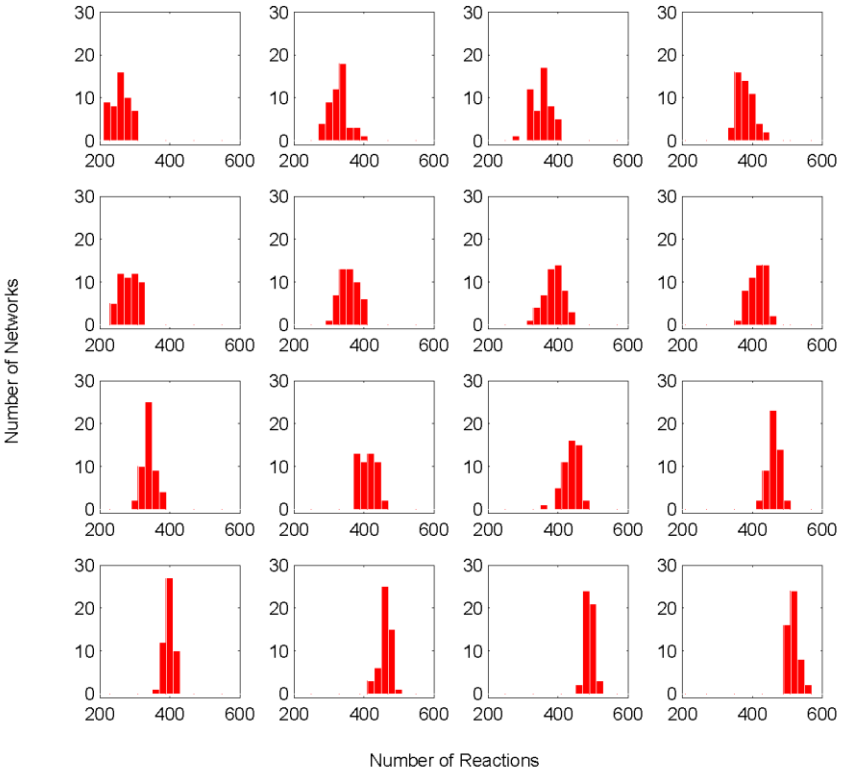

Figure S3. Distribution of number *R* of reactions. The distributions of R are shown for each combination of biosynthetic ability (*B*=20 first row, *B*=30 second row, *B*=40 third row, *B*=63 last row) and nutrient flexibility (*N*=20 first column, *N*=30 second column, *N*=40 third column, *N*=63 last column) that we examined, and for in total 800 minimal viable networks (50 Networks per panel).

Other Supplementary Material

Confidence Interval Calculation

We used the Matlab Statistical Package to calculate confidence intervals throughout the paper. The function makes use of the expression below to calculate 95% confidence limits for means:

$$P\left\{\bar{Y} - \frac{1.96\sigma}{\sqrt{n}} \leq \mu \leq \bar{Y} + \frac{1.96\sigma}{\sqrt{n}}\right\} = 1 - \alpha$$

where P is the probability that an actual mean μ lies in the indicated interval; $\bar{Y}$ is the sample mean, σ is the sample variance, and n is the sample size. α is 0.05, for we calculated 95% confidence intervals of mean.

To calculate 95% confidence limits for variances, we made use of the expression:

$$P\left\{\frac{(n-1)s^2}{X^2_{(\alpha/2)[n-1]}} \leq \sigma^2 \leq \frac{(n-1)s^2}{X^2_{\left(1-(\frac{\alpha}{2})\right)[n-1]}}\right\} = 1 - \alpha$$

Where P is the probability that the actual variance σ^2 lies in the indicated interval, where n is the sample size, s^2 is the sample variance, and X^2 is the value of the chi square distribution with n-1 degrees of freedom at a value $\alpha/2$ (for the left argument). α is 0.05, for we calculated 95% confidence intervals of variance.

Examples of reactions needed to metabolize new carbon sources

These examples take a network size reduction approach to illustrate the kinds of reactions needed to metabolize new carbon sources.

Example 1. We generated a minimal network that was required to synthesize *B*=63 biomass components with the ability to metabolize glucose, acetate and glycine. We used this network as a starting point for further reaction elimination to generate a minimal network that was required to be viable on acetate and glycine but not on glucose. The network able to grow only on acetate and glycine had two fewer reactions than the network able to grow on all three carbon sources. One of these is a glycolytic reaction (atp + fructose-6-phosphate → adp + fructose 1,6-bisphosphate + h), the other is an anaplerotic reaction (atp + oxaloacetate → adp + co2 + phosphoenolpyruvate).

Example 2: We generated a minimal network that was required to synthesize *B*=63 biomass components with the ability to metabolize glucose, glutamate and pyruvate. We used this network as a starting point for further reaction elimination to generate a minimal network that was required to be viable on glutamate and pyruvate but not on glucose. The network able to grow only on glutamate and pyruvate had two fewer reactions than the network able to grow on all three carbon sources. Both of them are glycolytic reactions ((1) atp + fructose-6-phosphate → adp + fructose 1,6-bisphosphate + h, (2) glucose-6-phosphate → fructose-6-phophate).

Examples of reactions needed to synthesize additional biomass molecules

These examples take a network size reduction approach to illustrate the kinds of reactions needed to synthesize additional biomass molecules.

Example 1: We generated a minimal network that was able to synthesize glutamate, asparagine and proline with glucose as the sole carbon source. We used this network as the starting point for further reaction elimination to generate a minimal network synthesizing glutamate and asparagine but, not proline. This resulted in the elimination of the following three reactions: (1) atp + glutamate → adp + glutamate 5-phosphate, (2) glutamate 5-phosphate + h + nadph → glutamate 5-semialdehyde + nadp + inorganic phosphate, (3) glutamate 5-semialdehyde → 1-pyrroline-5-carboxylate + h + h2o. These reactions are involved in the conversion of glutamate to proline.

Example 2: We generated a minimal network that was able to synthesize glutamate, arginine and adenonsine-3,5-bisphosphate in glucose. We used this network as the starting point for further reaction elimination to generate a minimal network synthesizing glutamate and arginine but not adenonsine-3,5-bisphosphate. This resulted in the elimination of the following three reactions: (1) acetyl-glutamate 5-semialdehyde + h2o → acetate + glutamate 5-semialdehyde, (2) acetyl-glutamate 5-semialdehyde --> 1-pyrroline-5-carboxylate + h + h2o, (3) 1-pyrroline-5-carboxylate + h + nadph → nadp + adenosine-3,5-bisphosphate. These reactions are involved in adenosine-3,5-bisphosphate synthesis.

3. Selection shapes the robustness of ligand-binding amino acids.

Tugce Bilgin [1,2], Isil Aksan Kurnaz [3], Andreas Wagner [1,2,4,5]

[1]Institute of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland, [2] The Swiss Institute of Bioinformatics, Lausanne, Switzerland, [3]Yeditepe University, Department of Genetics and Bioengineering, Istanbul, Turkey, [4] The Santa Fe Institute, Santa Fe, New Mexico, United States of America [5] Bioinformatics Institute, Agency for Science, Technology and Research (A*STAR), 30 Biopolis Street, Singapore 138671.

This chapter was published in Journal of Molecular Evolution, 76(5), pp 343-349 (doi:10.1007/s00239-013-9564-1)

3.1. Abstract

The phenotypes of biological systems are to some extent robust to genotypic changes. Such robustness exists on multiple levels of biological organization. We analyzed this robustness for two categories of amino acids in proteins. Specifically, we studied the codons of amino acids that bind or do not bind small molecular ligands. We asked to what extent codon changes caused by mutation or mistranslation may affect physicochemical amino acid properties or protein folding. We found that the codons of ligand-binding amino acids are on average more robust than those of non-binding amino acids. Because mistranslation is usually more frequent than mutation, we speculate that selection for error mitigation at the translational level stands behind this phenomenon. Our observations suggest that natural selection can affect the robustness of very small units of biological organization.

3.2. Introduction

Two computational approaches to characterize functionally important amino acids of a protein are widespread. The first focuses on the accessible surface area, which describes the accessibility of an amino acid by the solvent surrounding a protein (Lee and Richards 1971). Amino acids that are involved in binding ligands commonly occur in large and deep clefts on a protein's surface with low accessible surface area, which may help to increase the specificity and stability of binding (Bartlett et al. 2002; Laskowski et al. 1996). The analysis of solvent accessibility requires detailed knowledge of ligand binding sites, which is limited to proteins with known ligand-

bound structures. The second approach uses evolutionary conservation of amino acids (Capra et al. 2009; Lichtarge and Sowa 2002). For example, amino acids in catalytic sites of enzymes are more conserved on average (Bartlett et al. 2002). However, because evolutionary conservation is influenced by multiple factors, such as the divergence time between orthologs, the background rate of amino acid substitutions, and mutational biases (Sasidharan and Chothia 2007), information on conservation alone is not enough to characterize functional sites. Many studies thus combine these two approaches to improve the characterization of binding sites (Bartlett et al. 2002; Capra et al. 2009). Here we suggest a third, complementary approach that may help characterize specifically those amino acids that bind ligands. It focuses on their robustness to mutation or mistranslation. Because such amino acids are especially important for the function of a protein, they can be subject to selection increasing their robustness relative to non-ligand-binding amino acids.

Biological systems on multiple levels of organization are to some extent robust to genetic or environmental change. Examples include the genetic code of extant organisms, which is more robust to nucleotide changes than the vast majority of hypothetical alternative codes (Freeland and Hurst 1998); proteins, which can continue to function when many of their amino acids are mutated (Bowie et al. 1990; Guo et al. 2004; Huang et al. 1996; Loeb et al. 1989; Markiewicz et al. 1994; Suckow et al. 1996); gene regulatory circuits, whose phenotypes are to some extent robust to changes in regulatory interactions (Von Dassow and Odell 2002; Ingolia 2004; Isalan et al. 2008; Isalan et al. 2005; Li et al. 2006); and genome-scale metabolic networks, which can tolerate deletions of multiple enzyme-coding genes without detectable phenotypic effects in standard laboratory environments (Hillenmeyer et al. 2008).

Such robustness may reflect intrinsic system properties that may not have been shaped by natural selection. Alternatively, it may be the result of evolutionary adaptation, either to ameliorate the detrimental effects of DNA mutations, of environmental change, or of both.

Among the four principal ways in which random change in a codon can occur – DNA mutation, mistranscription, mRNA alteration, and mistranslation – we focus on mutation and the mistranslation of mRNA, which are well documented and probably most frequent. Such mistranslation occurs when a ribosome incorporates incorrect amino acids when synthesizing a protein from an mRNA template. There are at least three non-exclusive classes of evolutionary mechanisms by which the cost of mistranslation can be minimized. The first is selection of translational accuracy. Akashi (Akashi 1994) suggested that such selection causes genes or specific sites in genes to be encoded by codons that correspond to abundant tRNAs. Such high fidelity codons have higher chances of being accurately translated. The second is selection of translational robustness, which has been proposed by Drummond and Wilke (Drummond and Wilke 2008; Wilke and Drummond 2006). According to these authors, proteins (and especially highly expressed proteins) show evolved tolerance in their fold to missense translational errors (Zhou et al. 2009). The third involves error mitigation. Among those synonymous codons that encode the same amino acid, some are more robust to changes in individual nucleotides than others. That is, even though a random change in a robust codon may change the encoded amino acid, the new amino acid has, on average, similar physicochemical properties or does not perturb protein folding strongly (Archetti 2006; Archetti 2004a). In error mitigation, codons that are likely to be mistranslated into radically different amino acids are avoided.

Previous studies (Archetti 2004a; Najafabadi et al. 2007) showed that amino acids in eukaryotic and prokaryotic proteins are often encoded by codons whose mistranslation leads to the substitution of amino acids with limited deleterious effects. The most important of these three causes for our work is error mitigation.

Here we ask whether selection helps shape the codon usage of ligand-binding amino. To this end, we analyze the robustness of codons to mutation or mistranslation for two classes of codons in a protein, that is, codons that encode amino acids, which are or are not involved in the binding of a small molecular ligand. We use an estimator of robustness that incorporates the likely effects of an amino acid change on the physicochemical properties of an amino acid, and on protein folding. Our analysis shows that ligand-binding amino acids are on average more robust to mutation or mistranslation than non-binding amino acids, which is consistent with selection pressure for error mitigation.

3.3. Methods

We use a codon robustness score φ(c) derived from the weighted average load function of Ardell (Ardell 1998). This score aims to capture the predicted effect that a particular amino acid change has on the folding free energy of a protein and on physicochemical amino acid properties. Specifically,

$$\varphi(c) = \sum_{c'=1}^{9} p(c'|c)\, g[a(c), a(c')] \quad (1)$$

where summation is applied over all nine 1-mutant neighbors of a codon *c*. In this expression, *p(c'|c)* $\mathrm{p}(\mathrm{c}'|\mathrm{c})$ is the probability of changing a codon *c* for another codon *c'*, which is computed by multiplying the position-specific transition-transversion bias of mistranslations with the relative mistranslation frequency of a given nucleotide position. $g[a(c), a(c')]$ is the physicochemical effect or "cost" of substituting the amino acid encoded by codon *c*, *a(c)*, with that encoded by codon *c'*, *a(c').*We used the mutation matrix generated by Gilis and colleagues (Gilis et al. 2001), to calculate the cost of such an amino acid change (see Online Resource 1a for the matrix). This matrix uses information on changes in folding free energy and physiochemical properties of amino acid features, such as hydrophobicity, after an amino acid change. We note that the use of different substitution matrices would not strongly affect codon robustness scores (Najafabadi et al. 2005).

To give an example of how we calculated robustness scores, consider the codon *tta* that encodes leucine. It has nine 1-mutant neighbors, one of them being *tca*. Mistranslation from *tta* to *tca* corresponds to a transition at a nucleotide in the second position of a codon. With a relative mistranslation frequency of second position nucleotides of 0.1 and a transition transversion bias at second position nucleotides of 5, we computed a value of *p(tca|tta)* = 0.5, by multiply these numbers. Because *tca* encodes serine, we multiply this value by -1, which is the cost of mutation from leucine to serine based on the Mutation Matrix. We perform an exactly analogous calculation for all other 1-mutant neighbors of tta to arrive at $\varphi(tta)$. We then normalize this score by dividing it by the mean codon robustness of all leucine-encoding codons to eliminate the possible effects of amino acid biases. Finally, we

normalize the scores of all codons to the interval (0,1). The resulting scores are shown in Online Resource 2.

We used proteins in our analysis (Online Resource 3) that (i) have a reviewed (non-putative) 3D structure deposited in the protein data bank (PDB) (Bourne et al. 2004), (ii) exert their biological function as monomers, and (iii) bind to one of the small ligands in Online Resource 4. Binding to large molecules, such as other proteins, RNA and DNA involve highly divergent interaction types, and large interface areas, which might decrease functional importance of amino acids that contact a molecule (Lichtarge and Sowa 2002). As Clackson and Wells (Clackson and Wells 1995) showed, only a fraction of those residues actually contribute to binding. We therefore excluded those larger molecules. From this data set, we eliminated proteins that bind to multiple ligands, as well as proteins with more than 90 percent sequence identity to other proteins, thus arriving at a final data set of 275 proteins. We extracted a protein's coding exons by aligning the encoding gene (obtained from NCBI (Benson et al. 2004)) and the amino acid sequence (Bourne et al. 2004) with the tool Exonerate (Slater and Birney 2005).

3.4. Results and Discussion

We subdivided all amino acids of the proteins in our data set into two categories, those not involved in the binding of small ligands, and those involved in the binding of small ligands, which we defined to be lying within a 5 Ångstrom radius of a ligand

in the published tertiary structure. We then computed the robustness scores of codons using a wide range of mistranslation parameters. We varied two key parameters at each nucleotide position, the transition-transversion bias for which we used 5 different values between one and five, and the mistranslation frequency, for which we used 10 different values between 0.1 and 1. We then asked for each of 50 different parameter combinations whether robustness scores encoding the binding and the non- binding amino acids differ. We found that codons encoding ligand-binding amino acids are significantly more robust in all cases, with either the same or very similar *P* values (greatest $P < 10^{-30}$, smallest $P < 10^{-36}$, Wilcoxon Rank Sum test used throughout, unless otherwise mentioned). For the sake of simplicity, we thus used one particular parameter combination for all subsequent analyses, which is that of Freeland and Hurst (Freeland and Hurst 1998, see Online Resource 1b). Even though it may not be universally accurate (Kramer et al. 2010), it has also been employed by several other studies similar to ours and on a wide range of organisms (Archetti 2006; Archetti 2004; Drummond and Wilke 2008; Najafabadi et al. 2007). Figure 1a indicates the distribution of robustness scores based on these mistranslation biases ($P < 10^{-35}$, thick horizontal lines indicate medians).

We then repeated our analysis using an estimator of codon robustness by Archetti, which takes only the physicochemical effects of changed amino acids into account [15], but not the likely effect on protein misfolding, as does our estimator. Again, ligand-binding amino acids are significantly more robust when using this estimator ($P < 10^{-8}$). We also compared the *Z*-statistic, which is the standardized value of *U*, the Wilcoxon Ranked Sum statistic. For large samples like ours, *U* is normally distributed (Rice 1995), and thus Z follows a standard-normal ($N(0,1)$) distribution. We found

that the *Z*-statistic is much greater for our own robustness estimator ($Z = 12.38$) than for Archetti's estimator ($Z = 3.77$). This means that taking effects on misfolding into account, binding and non-binding amino acids differ to a much greater extent in their robustness.

We next compared differences in codon robustness to differences between more conventional indicators of functionally important binding amino acids. The first of them is the accessible surface area. We obtained the accessible surface area scores of each amino acid from (Kabsch and Sander 1983) and normalized them, so that they range between 0 and 1. As previous studies did (Bartlett et al. 2002; Laskowski et al. 1996), we found that ligand-binding amino acids indeed have significantly smaller accessible surface area ($P < 10^{-27}$) (Fig. 1b). The difference becomes more significant ($P < 10^{-300}$), when we remove the residues in the hydrophobic core, that is the residues with normalized accessible surface area values less than 0.25 from the analysis.

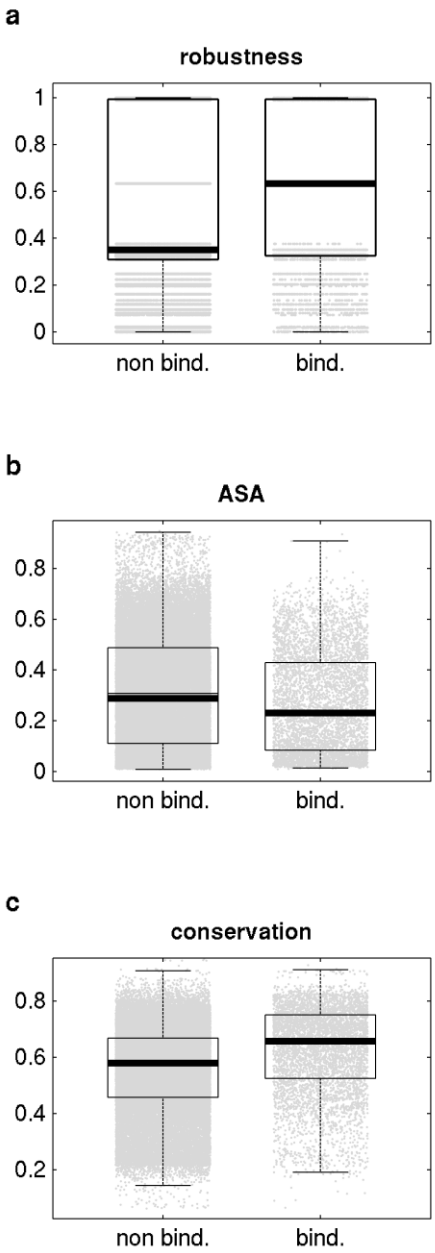

Fig. 1 Box-plot of **a)** robustness **b)** accessible surface area (ASA) **c)** amino acid conservation scores. Thick black horizontal lines in the middle of each box mark the median. The edges of the boxes correspond to the 25th and 75th percentiles. Data is based on a sample of n = 49,133 non-binding amino acids (left box in each panel) and of n = 5,552 ligand-binding amino acids (right box in each panel).

The second indicator is the extent of evolutionary conservation. We compared amino acid conservation scores (obtained from (Goldenberg et al. 2009)) for ligand-binding and non-binding amino acids in our data set. In line with previous studies (Bartlett et al. 2002; Capra et al. 2009), we found that the binding amino acids are significantly more conserved ($P < 10^{-258}$) (Fig. 1c). We next asked whether codon robustness

discriminates to a similar extent between binding and non-binding amino acids as do these two quantities. To this end, we examined again the Z-statistic of the Wilcoxon test, and found that evolutionary conservation differs most between binding and non-binding amino acids (Z = 34.35), the accessible surface area differs least (Z = 10.89) and codon robustness lies in between them (Z = 12.38). These observations suggest that robustness, while not as informative as evolutionary conservation, may have similar value as accessible surface area to characterize functionally important amino acids. Finally, we calculated the association of codon robustness with the other two indicators, and found that neither accessible surface area (r^2 = 0.15, P <10^{-300}) nor evolutionary conservation (r^2 = -0.06, P <10^{-48}) are strongly correlated with robustness. These weak correlations suggest that robustness is complementary to the two other two quantities in characterizing ligand-binding amino acids.

We next asked whether different genes also differ in the robustness scores of the ligand-binding amino acids they encode. To this end, we ranked proteins according to the mean robustness score and displayed the corresponding data as a rank plot (Fig. 2a). Specifically, the plot shows the rank-ordered codon robustness scores of ligand-binding amino acids (orange dots), together with the robustness of the non-binding amino acids (grey dots). For 63 percent or 173 proteins, the mean codon robustness of the binding amino acids was greater than the mean codon robustness of the non-binding amino acids, where the mean is taken over all non-binding amino acids in all proteins. A minority of 37 percent of proteins had a lower robustness of binding amino acids than those of non-binding amino acids. To investigate these differences in robustness further, we focused on two classes of proteins, the 28 proteins in the bottom 10^{th}-percentile (blue shading in Fig. 2a), and the 28 proteins in the top 10^{th}-

percentile (pink shading in Fig. 2a). We refer to them as the proteins with the lowest and highest robustness of ligand-binding amino acids.

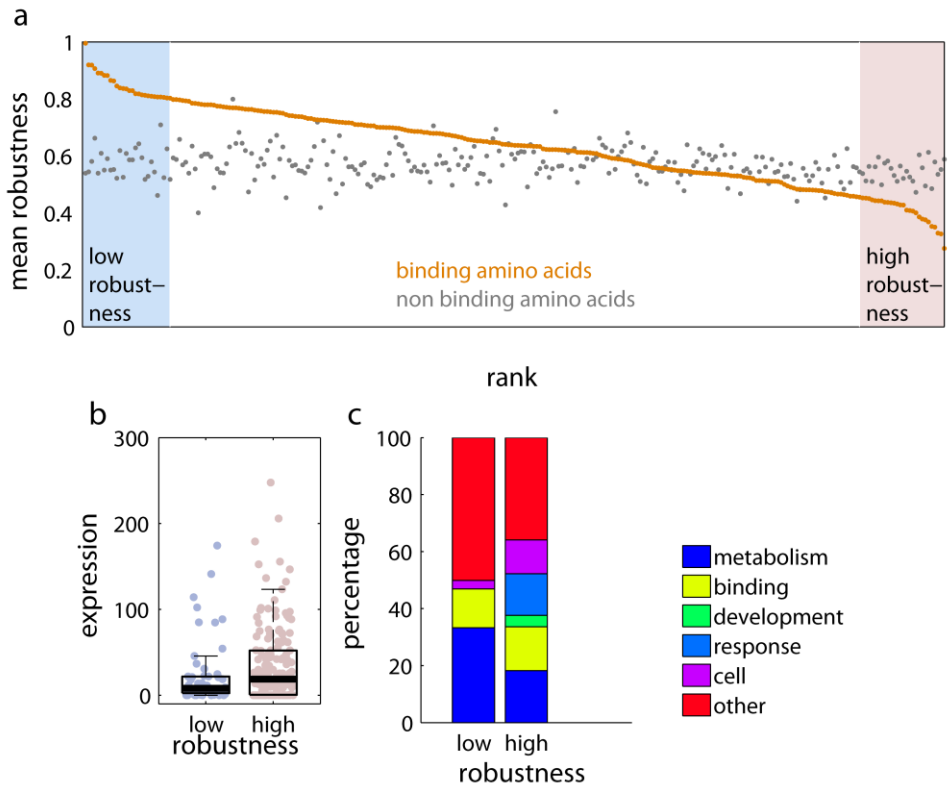

Fig. 2 a) Plot of mean robustness scores, ranked based on the mean robustness of ligand-binding amino acids for each of our 275 study proteins. Grey dots correspond to mean robustness scores of non-binding amino acids, orange dots correspond to mean robustness scores of ligand binding amino acids. The pink and blue regions correspond to the upper and lower 10^{th} percentiles in robustness, respectively. **b)** Box-plot of RNA expression levels in the brain for the proteins with lowest (right box) and highest (left box) robustness of ligand-binding amino acids. The edges of the boxes correspond to the 25th and 75th percentiles. Data is based on a sample of gene expression values n = 168 for the left box, n = 56 for the right box. **c)** Gene ontology (Ashburner et al. 2000) functional annotations of proteins with lowest (right bar) and highest (left bar) robustness of ligand-binding amino acids. Legend displays the major functional classes presented in the bars.

We first asked whether the genes encoding these two classes of proteins differ substantially in their expression. To this end, we used a gene expression data set

(Brawand et al. 2011) from 6 humans and five different organs (brain, heart, kidney, liver and testis), obtained through high throughput RNA sequencing (RNA seq). The genes in the highest robustness category did not show significantly higher expression when we analyzed pooled data from all organs, nor when we analyzed data from four of the five organs. The only exception was expression data from the brain, where these proteins were significantly more highly expressed (P = 0.029) (Fig. 2b). Although the signal becomes insignificant after a correction for false discovery rate (FDR, Benjamini and Hochberg 1995)), this pattern is consistent with an earlier analysis (Drummond and Wilke 2008), which showed that selection for translationally robust codons is strongest in brain and other neural tissues. The likely reason is the extreme sensitivity of neuronal functions to protein misfolding and dysfunction, which is associated with neurodegenerative diseases and neurotoxic effects (Bucciantini et al. 2002; Lee et al. 2006b). Using the Gene Ontology (Ashburner et al. 2000) classification of gene functions, we also found that proteins with highest robustness are significantly more enriched in functions related to development, differentiation (Exact Binomial Test, $P < 10^{-3}$), whereas proteins with lowest robustness are significantly more enriched in metabolic functions (Fig. 2c). In sum, these analyses reveal differences between proteins whose ligand-binding amino acids differ most in their robustness, although they fall short of explaining the low robustness we observe for these amino acids in some proteins.

That functionally or structurally important amino acids or codons are subject to special constraints has been proposed by previous work and in contexts different from ours. First, Bartlett and colleagues provided evidence that catalytic sites harbor certain classes of amino acids. Specifically, charged amino acids are more often found in

catalytic sites, whereas hydrophobic amino acids are more often found in the structure-stabilizing hydrophobic core (Bartlett et al. 2002). Second, Pakula and Sauer (Pakula and Sauer 1989) showed that such sites are highly constrained in the substitutions they can tolerate. Third, Zhou and colleagues (Zhou et al. 2009) provided evidence in several eukaryotes and prokaryotes that some parts of proteins are more sensitive to misfolding, and show a more constrained codon usage, the phenomenon that different synonymous codons for the same amino acid are not used equally frequently in protein coding genes (Akashi 1994; Akashi and Eyre-Walker 1998; Comeron and Aguadé 1998; Duret 2002; Gouy and Gautier 1982; Ikemura 1985; Ikemura 1981; Moriyama and Hartl 1993; Plotkin et al. 2004; Sharp et al. 1986; Sharp and Li 1987; Stoletzki and Eyre-Walker 2007). In general, the strength of this bias varies within genes and becomes stronger at functionally important sites (Akashi 1994; Stoletzki and Eyre-Walker 2007). In sum, our observations that robust codons are favored at ligand-binding amino acids are consistent with a broad range of related evidence.

Limitations of our analysis include the moderate number of 275 proteins we could study, as well as a small number of binding amino acids (18 on average) per protein, which renders rigorous statistical analysis of individual proteins infeasible. Despite these limitations, our joint analysis of multiple proteins showed a significant preference of robust codons in ligand-binding pockets of proteins, exactly where amino acid changes can have a highly detrimental effect on protein function.

Another limitation is that codon robustness *alone* – like accessible surface area and conservation – does not have much power to predict ligand-binding sites. To predict

such sites, more complex models incorporating multiple characterizing elements are necessary (see for example, Capra et al. 2009; Lichtarge and Sowa 2002; Wass et al. 2011). Because codon robustness differs more than accessible surface area between ligand binding and non-binding amino acids, our approach can help improve such models and their predictive power.

Selection may have favored robust codons in ligand-binding amino acids because they are robust to mutation or to translation. Although mistranslation is not genetic change –it leaves the DNA encoding a mRNA unchanged – it does alter the encoded protein randomly (Drummond and Wilke 2009). Translational error rates in microbes have been estimated at 10^{-3}-10^{-4} per codon. This number is at least five orders of magnitude higher than typical mutation rates (Kramer et al. 2010; Kramer and Farabaugh 2007; Ogle and Ramakrishnan 2005b). At this error rate, 15 percent of protein molecules would contain at least one mistranslated amino acid. Translation errors can induce protein misfolding, aggregation, toxicity and cell death, which underlie a broad array of neurodegenerative diseases (Bucciantini et al. 2002; Lee et al. 2006b). Also, mistranslation at functionally important sites can disrupt protein function (Guo et al. 2004; Markiewicz et al. 1994). For these reasons, we speculate that selection for error mitigation at the translational level is the prevalent driving force of high robustness in codons that encode ligand-binding amino acids. Why a minority of ligand-binding amino acids has especially low codon robustness remains an open question for future work.

3.5. Acknowledgements

AW and TB acknowledge support through Swiss National Science Foundation grants 315230-129708, as well as through the YeastX project of SystemsX.ch, and the University Priority Research Program in Systems Biology at the University of Zurich. TB acknowledges support through TUBITAK BIDEB 2209 and thanks Turkan Haliloglu and Niv Sabath for helpful discussions.

3.6. References

Akashi, H. (1994). Synonymous Codon Usage in Drosophila Melanogaster: Natural Selection and Translational Accuracy. *Genetics*, *136*(3), 927–935. Retrieved from http://www.genetics.org/cgi/content/abstract/136/3/927

Akashi, H., & Eyre-Walker, A. (1998). Translational selection and molecular evolution. *Current Opinion in Genetics & Development*, *8*(6), 688–93. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9914211

Archetti, M. (2004). Selection on codon usage for error minimization at the protein level. *Journal of Molecular Evolution*, *59*(3), 400–15. doi:10.1007/s00239-004-2634-7

Archetti, M. (2006). Genetic robustness and selection at the protein level for synonymous codons. *Journal of Evolutionary Biology*, *19*(2), 353–65. doi:10.1111/j.1420-9101.2005.01029.x

Ardell, D. H. (1998). On error minimization in a sequential origin of the standard genetic code. *Journal of Molecular Evolution*, *47*(1), 1–13. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9664691

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., … Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29. doi:10.1038/75556

Bartlett, G. J., Porter, C. T., Borkakoti, N., & Thornton, J. M. (2002). Analysis of catalytic residues in enzyme active sites. *Journal of Molecular Biology*, *324*(1), 105–121. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0022283602010367

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B Methodological*, *57*(1), 289–300. doi:10.2307/2346101

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Wheeler, D. L. (2004). GenBank: update. *Nucleic Acids Research*, *32*(Database issue), D23–D26. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=308779&tool=pmcentrez&rendertype=abstract

Bourne, P. E., Addess, K. J., Bluhm, W. F., Chen, L., Deshpande, N., Feng, Z., … Berman, H. M. (2004). The distribution and query systems of the RCSB Protein Data Bank. *Nucleic Acids Research*, *32*(Database issue), D223–D225. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=308830&tool=pmcentrez&rendertype=abstract

Bowie, J. U., Reidhaar-Olson, J. F., Lim, W. A., & Sauer, R. T. (1990). Deciphering the message in protein sequences: tolerance to amino acid substitutions. *Science*, *247*(4948), 1306–1310. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2315699

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., … Kaessmann, H. (2011). The evolution of gene expression levels in mammalian organs. *Nature*, *478*(7369), 343–348. doi:10.1038/nature10532

Bucciantini, M., Giannoni, E., Chiti, F., Baroni, F., Formigli, L., Zurdo, J., … Stefani, M. (2002). Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *Nature*, *416*(6880), 507–511. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11932737

Capra, J. A., Laskowski, R. A., Thornton, J. M., Singh, M., & Funkhouser, T. A. (2009). Predicting Protein Ligand Binding Sites by Combining Evolutionary Sequence Conservation and 3D Structure. *PLoS Computational Biology*, *5*(12), 18. Retrieved from http://dx.doi.org/10.1371/journal.pcbi.1000585

Clackson, T., & Wells, J. A. (1995). A hot spot of binding energy in a hormone-receptor interface. *Science*, *267*(5196), 383–386. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7529940

Comeron, J. M., & Aguadé, M. (1998). An evaluation of measures of synonymous codon usage bias. *Journal of Molecular Evolution*, *47*(3), 268–74. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9732453

Drummond, D. A., & Wilke, C. O. (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*, *134*(2), 341–352. doi:10.1016/j.cell.2008.05.042

Drummond, D. A., & Wilke, C. O. (2009). The evolutionary consequences of erroneous protein synthesis. *Nature Reviews Genetics*, *10*(10), 715–724. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19763154

Duret, L. (2002). Evolution of synonymous codon usage in metazoans. *Current Opinion in Genetics & Development*, *12*(6), 640–649. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12433576

Freeland, S. J., & Hurst, L. D. (1998). The genetic code is one in a million. *Journal of Molecular Evolution*, *47*(3), 238–48. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9732450

Gilis, D., Massar, S., Cerf, N. J., & Rooman, M. (2001). Optimality of the genetic code with respect to protein stability and amino-acid frequencies. *Genome Biology*, *2*(11), RESEARCH0049. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=60310&tool=pmcentrez&rendertype=abstract

Goldenberg, O., Erez, E., Nimrod, G., & Ben-Tal, N. (2009). The ConSurf-DB: pre-calculated evolutionary conservation profiles of protein structures. *Nucleic Acids Research*, *37*(Database issue), D323–D327. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2686473&tool=pmcentrez&rendertype=abstract

Gouy, M., & Gautier, C. (1982). Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Research*, *10*(22), 7055–7074. doi:10.1093/nar/gkn1031

Guo, H. H., Choe, J., & Loeb, L. A. (2004). Protein tolerance to random amino acid change. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(25), 9205–9210. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=438954&tool=pmcentrez&rendertype=abstract

Hillenmeyer, M. E., Fung, E., Wildenhain, J., Pierce, S. E., Hoon, S., Lee, W., … Giaever, G. (2008). The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science (New York, N.Y.)*, *320*(5874), 362–5. doi:10.1126/science.1150021

Huang, W., Petrosino, J., Hirsch, M., Shenkin, P. S., & Palzkill, T. (1996). Amino acid sequence determinants of beta-lactamase structure and activity. *Journal of Molecular Biology*, *258*(4), 688–703. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8637002

Ikemura, T. (1981). Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the

E. coli translational system. *Journal of Molecular Biology*, *151*(3), 389–409. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/6175758

Ikemura, T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. *Molecular Biology and Evolution*, *2*(1), 13–34.

Ingolia, N. T. (2004). Topology and Robustness in the Drosophila Segment Polarity Network. *PLoS Biology*, *2*(6), e123. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15208707

Isalan, M., Lemerle, C., Michalodimitrakis, K., Horn, C., Beltrao, P., Raineri, E., … Serrano, L. (2008). Evolvability and hierarchy in rewired bacterial gene networks. *Nature*, *452*(7189), 840–845. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18421347

Isalan, M., Lemerle, C., & Serrano, L. (2005). Engineering Gene Networks to Emulate Drosophila Embryonic Pattern Formation. *PLoS Biology*, *3*(3), e64. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15736977

Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, *22*(12), 2577–2637. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/6667333

Kramer, E. B., & Farabaugh, P. J. (2007). The frequency of translational misreading errors in E. coli is largely determined by tRNA competition. *Rna New York Ny*, *13*(1), 87–96. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17095544

Kramer, E. B., Vallabhaneni, H., Mayer, L. M., & Farabaugh, P. J. (2010). A comprehensive analysis of translational missense errors in the yeast Saccharomyces cerevisiae. *Rna New York Ny*, *16*(9), 1797–1808. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20652078

Laskowski, R. A., Luscombe, N. M., Swindells, M. B., & Thornton, J. M. (1996). Protein clefts in molecular recognition and function. *Protein Science*, *5*(12), 2438–2452. Retrieved from http://discovery.ucl.ac.uk/1344567/

Lee, B., & Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *Journal of Molecular Biology*, *55*(3), 379–400. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/5551392

Lee, J. W., Beebe, K., Nangle, L. A., Jang, J., Longo-Guess, C. M., Cook, S. A., … Ackerman, S. L. (2006). Editing-defective tRNA synthetase causes protein misfolding and neurodegeneration. *Nature*, *443*(7107), 50–55. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16906134

Li, S., Assmann, S. M., & Albert, R. (2006). Predicting Essential Components of Signal Transduction Networks: A Dynamic Model of Guard Cell Abscisic Acid Signaling. *PLoS Biology*, *4*(10), 17. Retrieved from http://arxiv.org/abs/q-bio/0610012

Lichtarge, O., & Sowa, M. E. (2002). Evolutionary predictions of binding surfaces and interactions. *Current Opinion in Structural Biology*, *12*(1), 21–27. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11839485

Loeb, D. D., Swanstrom, R., Everitt, L., Manchester, M., Stamper, S. E., & Hutchison, C. A. (1989). Complete mutagenesis of the HIV-1 protease. *Nature*, *340*(6232), 397–400. Retrieved from http://apps.isiknowledge.com/full_record.do?product=WOS&search_mode=GeneralSearch&qid=68&SID=S2LepGod4l@AMamB@e1&page=1&doc=1

Markiewicz, P., Kleina, L. G., Cruz, C., Ehret, S., & Miller, J. H. (1994). Genetic studies of the lac repressor. XIV. Analysis of 4000 altered Escherichia coli lac repressors reveals essential and non-essential residues, as well as “spacers” which do not require a specific sequence. *Journal of Molecular Biology*, *240*(5), 421–433. Retrieved from http://dx.doi.org/10.1006/jmbi.1994.1458

Moriyama, E. N., & Hartl, D. L. (1993). Codon usage bias and base composition of nuclear genes in Drosophila. *Genetics*, *134*(3), 847–858. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1205521&tool=pmcentrez&rendertype=abstract

Najafabadi, H. S., Goodarzi, H., & Torabi, N. (2005). Optimality of codon usage in Escherichia coli due to load minimization. *Journal of Theoretical Biology*, *237*(2), 203–209. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15932760

Najafabadi, H. S., Lehmann, J., & Omidi, M. (2007). Error minimization explains the codon usage of highly expressed genes in Escherichia coli. *Gene*, *387*(1-2), 150–155. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17097242

Ogle, J. M., & Ramakrishnan, V. (2005). Structural insights into translational fidelity. *Annual Review of Biochemistry*, *74*(1), 129–177. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15952884

Pakula, A. A., & Sauer, R. T. (1989). Genetic analysis of protein stability and function. *Annual Review of Genetics*, *23*, 289–310. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2694933

Plotkin, J. B., Robins, H., & Levine, A. J. (2004). Tissue-specific codon usage and the expression of human genes. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(34), 12588–12591. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=515101&tool=pmcentrez&rendertype=abstract

Rice, J. A. (1995). *Mathematical Statistics and Data Analysis*. *Higher Education* (Vol. 72, p. 330). Duxbury Press. doi:10.2307/3619963

Sasidharan, R., & Chothia, C. (2007). The selection of acceptable protein mutations. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(24), 10080–10085. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17540730

Sharp, P. M., & Li, W. H. (1987). The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research*, *15*(3), 1281–1295. Retrieved from http://nar.oxfordjournals.org/cgi/content/abstract/15/3/1281

Sharp, P. M., Tuohy, T. M., & Mosurski, K. R. (1986). Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Research*, *14*(13), 5125–5143. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=311530&tool=pmcentrez&rendertype=abstract

Slater, G. S. C., & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, *6*(1), 31. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15713233

Stoletzki, N., & Eyre-Walker, A. (2007). Synonymous codon usage in Escherichia coli: selection for translational accuracy. *Molecular Biology and Evolution*, *24*(2), 374–81. doi:10.1093/molbev/msl166

Suckow, J., Markiewicz, P., Kleina, L. G., Miller, J., Kisters-Woike, B., & Müller-Hill, B. (1996). Genetic studies of the Lac repressor. XV: 4000 single amino acid substitutions and analysis of the resulting phenotypes on the basis of the protein structure. *Journal of Molecular Biology*, *261*(4), 509–523. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8794873

Von Dassow, G., & Odell, G. M. (2002). Design and constraints of the Drosophila segment polarity module: robust spatial patterning emerges from intertwined cell state switches. *The Journal of Experimental Zoology*, *294*(3), 179–215. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12362429

Wass, M. N., David, A., & Sternberg, M. J. E. (2011). Challenges for the prediction of macromolecular interactions. *Current Opinion in Structural Biology*, *21*(3), 382–390. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21497504

Wilke, C. O., & Drummond, D. A. (2006). Population Genetics of Translational Robustness. *Genetics*, *173*(1), 473–481. Retrieved from http://arxiv.org/abs/q-bio/0509031

Zhou, T., Weems, M., & Wilke, C. O. (2009). Translationally optimal codons associate with structurally sensitive sites in proteins. *Molecular Biology and Evolution*, *26*(7), 1571–80. doi:10.1093/molbev/msp070

3.7. Supplementary Material

Online Resource 1

	A																			
A	+7	**C**																		
C	−3	+7	**D**																	
D	0	−4	+7	**E**																
E	0	−5	+2	+7	**F**															
F	0	−2	−2	−3	+7	**G**														
G	−2	−4	−1	−2	−3	+7	**H**													
H	+1	−2	+1	+1	−1	0	+7	**I**												
I	−2	−3	−4	−4	0	−4	−2	+7	**K**											
K	+1	−5	+2	+3	−2	−1	+1	−3	+7	**L**										
L	−1	−3	−3	−3	0	−4	−2	0	−2	+7	**M**									
M	−1	−3	−1	−1	0	−2	0	0	−1	+1	+7	**N**								
N	0	−3	+2	+2	−2	0	+2	−3	+2	−2	−1	+7	**P**							
P	−3	−6	−1	−2	−4	−4	−2	−5	−1	−5	−4	−1	+7	**Q**						
Q	+1	−4	+2	+3	−1	−1	+2	−2	+3	−1	0	+2	−2	+7	**R**					
R	+1	−4	+2	+2	−2	−1	+1	−3	+3	−2	0	+2	−2	+2	+7	**S**				
S	+1	−2	+2	+1	−1	0	+2	−2	+2	−1	0	+2	−1	+2	+2	+7	**T**			
T	0	−2	+1	0	0	−1	+2	−1	+1	−1	0	+1	−1	+1	+1	+2	+7	**V**		
V	−1	−3	−3	−4	0	−4	−2	+1	−3	0	0	−3	−4	−2	−2	−1	0	+7	**W**	
W	+1	−2	0	−1	0	−2	+1	−1	0	−1	+1	0	−3	0	0	+1	0	−1	+7	**Y**
Y	0	−1	−1	−2	+1	−1	+1	−1	−1	−1	0	0	−3	0	−1	+1	+1	0	+1	+7

Online Resource 1a The Gilis Mutation Matrix (Gilis et al. 2001). Mutations to stop codons have the value -10.

	First Nucleotide	Second Nucleotide	Third Nucleotide
Relative mistranslation frequency	0.5	0.1	1
Transition-transversion bias of mistranslations	2	5	1

Online Resource 1b Relative frequencies and biases of mistranslation used in the robustness score calculations (Freeland and Hurst 1998).

Online Resource 2

	T	**C**	**A**	**G**	
T	0.798	0.955	0.606	0.585	**T**
	0.798	0.955	0.606	0.585	**C**
	0.857	0.944	0.326	0.000	**A**
	0.870	0.948	0.293	0.407	**G**
C	0.923	0.875	0.877	0.906	**T**

	0.923	0.875	0.877	0.906	**C**
	0.990	0.875	0.803	0.898	**A**
	1.000	0.875	0.803	0.994	**G**
A	0.873	0.943	0.880	0.842	**T**
	0.873	0.943	0.880	0.842	**C**
	0.805	0.943	0.805	0.805	**A**
	0.684	0.944	0.806	0.902	**G**
G	0.939	0.912	0.854	0.882	**T**
	0.939	0.912	0.854	0.882	**C**
	0.939	0.912	0.794	0.813	**A**
	0.929	0.912	0.794	0.899	**G**

Online Resource 2 Codon robustness scores for all 64 codons.

Online Resource 3

1A17	1I8U	1PB5	1WEM	2CO8	2EGM	2PJZ
1A44	1IG5	1PHR	1WEO	2CON	2EGP	2PL5
1A76	1IGS	1PLQ	1WES	2COT	2EJ4	2PLZ
1A77	1IGV	1PNT	1WEV	2CR8	2EKL	2PRD
1AIN	1ILE	1PXE	1WFH	2CS3	2ELI	2PT1
1ALA	1INP	1Q1F	1WFP	2CSV	2EWT	2Q18
1ANN	1IQ3	1Q1N	1WG2	2CSY	2FC6	2UY2
1AVC	1ISP	1QGO	1WIL	2CT0	2FC7	2UY9
1AXN	1J55	1QME	1WKB	2CT2	2FGF	2V07
1B5M	1JE6	1QWG	1WY9	2CT5	2FR7	2V5D
1BCI	1JNI	1R03	1X0T	2CT7	2GJL	2VRG
1BDB	1K49	1R79	1X3H	2CU2	2GQJ	2VW0
1BOR	1KIT	1RDV	1X4J	2CUL	2H0L	2W38
1C6S	1KPF	1RK9	1X4V	2CVC	2H3M	2YQL
1CO4	1KT0	1RQG	1X4W	2CW7	2H6E	2YQM
1CTL	1L4B	1RWJ	1X5W	2D8S	2H9L	2YQP
1D9V	1LC0	1SBP	1X61	2D8T	2HBK	2YS2
1DHR	1LC5	1SIQ	1X62	2D8U	2HPJ	2YT5
1DT1	1M36	1SQ9	1X64	2D9G	2HQQ	2YUU
1DXH	1M47	1SRK	1X6E	2D9K	2I0M	2YYG
1EE9	1M6N	1TFA	1XF7	2DAR	2I4I	2Z0Z
1EF4	1M9I	1TFI	1XFI	2DB6	2I5O	2ZAO
1EK5	1M9O	1U96	1XWI	2DJ7	2I6J	2ZJ2
1F62	1MDF	1UHN	1Y02	2DJA	2ILK	2ZLB
1FF9	1MGT	1UKY	1YGT	2DJB	2J8H	3B4X
1FLV	1MHO	1UL4	1YK4	2DJR	2JGU	3BI9
1FUE	1MNH	1UM8	1YWQ	2DUL	2JM4	3C5K
1G0W	1MPT	1UPQ	1ZMR	2E5S	2JMI	3CJ7
1G25	1MVL	1UV0	1ZW8	2E61	2JMO	3CU4
1G2P	1N67	1UZ5	2A3M	2E6I	2JQ6	3DD4
1GCG	1NJ3	1V5N	2A4E	2E6R	2JYD	3EIE

1GGZ	1NMW	1V9M	2A6Y	2E6S	2K1W	3F6Y
1GV9	1NOX	1VCI	2AEU	2E73	2K4X	3H2X
1HDR	1NSJ	1VCN	2AYJ	2EBL	2KGG	4CLN
1HEX	1NVG	1VD6	2B1O	2ECI	2MM1	4TNC
1HI5	1NX2	1VJI	2B71	2ECL	2NVH	
1HQV	1NZ3	1W5D	2BWK	2ECM	2O72	
1HTN	1OG3	1WAB	2BWL	2ECN	2OO3	
1HYJ	1OL6	1WD2	2CCQ	2ECT	2P5Y	
1I7P	1OOT	1WEE	2CEI	2ECY	2PCN	

Online Resource 3 List of PDB identifiers for all proteins used in this study.

Online Resource 4

acetate	chlorine	iron	potassium
adenosine diphosphate	copper	magnesium	protoporphyrin
adenosine monophosphate	flavin adenin dinucleotide	manganese	sodium
adenosine triphosphate	flavin mononucleotide	mercury	sulfate
bromide	glycerol	nicotinamide adenine dinucleotide	zinc
calcium	heme c	phophate	

Online Resource 4 List of ligands used in this study in alphabetic order.

4. Tandem repeats and increased expression divergence in primate genes

Tugce Bilgin [1, 2], Mark D. Robinson [3], Andreas Wagner [1, 2, 4]

[1]Institute of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland, [2] The Swiss Institute of Bioinformatics, Lausanne, Switzerland, [3] Institute of Molecular Life Sciences, University of Zurich, 8057 Zurich, Switzerland, [4] The Santa Fe Institute, Santa Fe, New Mexico, United States of America

4.1. Abstract

Tandem repeats tend to be highly variable in their length, and are thus an important source of genetic variation. Repeat variants are associated with many diseases, including cancers, but they may also play a role in the evolution of gene regulation. We here analyze the potential influence of tandem repeats on gene expression evolution in the genomes of humans, chimpanzees, and macaques. To this end, we identified tandem repeats with repeat units of up to 50 base pairs. We found that 30 percent of 13,035 orthologous genes in the three species contained tandem repeats within five kilo base pairs (kbp) of their transcription start site. Genes with repeat-containing promoters show significantly higher expression divergence in all three species. Moreover, duplicate genes show greater expression divergence if one or both duplicates contain repeats in their promoter. Genes with repeats in their 3' untranslated region, in introns, and in exons also show higher expression divergence. Hence, tandem repeats, far from just being a source of genetic diseases, may contribute substantially to the divergence of gene expression in primates.

4.2. Introduction

Gene expression divergence is a major driver of phenotypic change in evolution (Carroll 2000; Dixon et al. 2007; Jordan et al. 2005; King and Wilson 1975; Li et al. 2010; Ponting 2008; Wray et al. 2003a). A key challenge is to understand the factors contributing to this divergence. The most prominent such factor is single-nucleotide

polymorphisms (SNPs), which has long been known to associate with phenotypic variation (Stranger et al. 2007a; Stranger et al. 2007b; Stranger et al. 2005). SNPs in human cis-regulatory regions can explain more than 70 per cent of gene expression variation (Li et al. 2010; Rockman and Wray 2002; Stranger et al. 2007a; Stranger et al. 2007b). A second potentially important factor is copy-number variation, which comprises polymorphic duplications of chromosome segments, and correlates with gene expression levels (Hurles et al. 2008; Stranger et al. 2007a).

Tandem repeats are candidates for a third major contributor to gene expression divergence (Gemayel et al. 2010; Kashi and King 2006; Rockman and Wray 2002; Vinces et al. 2009). These are DNA tracts in which a short (1-50) base-pair motif, the repeat unit, is repeated several times in tandem, i.e., in a closely spaced, head-to-tail orientation. They are among the most variable loci in the human genome, experiencing mutations in the copy number of repeat units that are 100 to 100,000 times more frequent than point mutations, and occur at a rate of 10^{-3} to 10^{-7} copy number alterations per cell division (Brinkmann et al. 1998; Legendre et al. 2007; Li et al. 2002; Weber and Wong 1993). Most mutations giving rise to variable tandem repeats result from replication slippage (Levinson and Gutman 1987; Li et al. 2002; Schlötterer 2000; Schlotterer and Tautz 1992; Webster et al. 2002).

While variable repeats found in genes are responsible for over 40 human neurological/neuromuscular diseases, such as Huntington's disease and the spinocerebellar ataxias (reviewed in (Pearson et al. 2005a) and in (Bates 1996)), not all such repeats cause disease. They can also contribute to normal phenotypic variation and help fine-tune gene products and their gene expression adaptively

(Fondon et al. 2008; Kashi and King 2006). For example, the expression of *H. influenzae* fimbriae is subject to reversible phase variation between three expression levels, which is required for different stages of infection. Responsible for this variation are copy number changes of repetitive TA tracks in *hifA* and *hifB* promoter regions (van Ham et al. 1993).

Repeat-induced gene expression variation is not restricted to microbes, but can also induce non-pathological variation in higher organisms. In tilapia (*Oreochromis niloticus*), an important aquacultural fish, variable CA repeats in the promoter of the *prl1* gene, which is involved in osmoregulation, are associated with both gene expression levels and gene functionality (Streelman and Kocher 2002). An especially remarkable example regards features of a dog's snout, such as the degree of dorsoventral nose bend and midface length, which correlate with the ratio of the length of two tandem repeats in a gene that regulates bone formation (Fondon and Garner 2004). Repeat polymorphisms may also contribute to behavioral and cognitive functions. For example, the presence of tandem repeats in the 5' untranslated region of the vole *vasopressin 1a receptor* gene correlates with social behavior (Fondon et al. 2008). Mutations in this gene have been implicated in autism (Kim et al. 2002). The gene's orthologue in chimpanzees has a partial deletion in those tandem repeats, whereas the more social human and bonobo (Waal 2009) have the complete set of repeats (Hammock and Young 2005).

Because gene expression changes are so important in evolution, it is important to study mechanisms that can permit rapid expression change on short evolutionary time scales (Choi and Kim 2008; Landry et al. 2007; Tirosh et al. 2009; Tirosh et al. 2006;

Wray et al. 2003a). Promoter features such as TATA boxes, nucleosome-covering, as well as tracts of tandem repeats can mediate such change (Tirosh him et al., 2009). Among these features, tandem repeats are especially attractive study objects, because of their high incidence in regulatory regions (Gemayel et al. 2010; Payseur et al. 2011) and their high mutability (Brinkmann et al. 1998; Legendre et al. 2007; Li et al. 2002; Weber and Wong 1993).

One species where an association between tandem repeat variation and gene expression divergence has been demonstrated, and where regulatory regions are enriched in repeats is the yeast *Saccharomyces cerevisiae* (Vinces et al. 2009). Proximal upstream sequences of human repeat containing genes, which are likely to function as promoters, are also enriched for repeats compared to distant upstream and coding sequences (Supplementary Material in Vinces et al., 2009). This observation suggests a similar role for repeats in human expression evolution. To characterize this role, we examined the influence of tandem repeats in promoters on gene expression divergence in humans, chimpanzees, and macaques. We found that the presence of tandem repeats is indeed associated with greater expression divergence. This association holds for all three species, and for genes expressed in each of several organs. These observations extend to repeats in the promoters of duplicate genes, and to repeats in other gene regions that can influence gene regulation, including 3' untranslated regions, introns, and exons.

4.3. Results

Many primate promoters contain tandem repeats.

We identified genes with tandem repeats in the 5 kilo base pairs (kbp) upstream from the transcription start site of *n* = 13,035 orthologous genes in humans, chimpanzees, and macaques. We found that, depending on the species, on average 29-31% of all genes in these orthologs harbored tandem repeats. Specifically, 3820, 3910 and 4032 gene promoters harbored at least one repeat in human, chimpanzee and macaque, respectively.

Next we wanted to find out whether repeats contained in the first 5 kbp base pairs upstream of a gene are relevant for gene expression. To this end, we retrieved data on DNase hypersensitive site locations from ENCODE (Material et al. 2004) using the UCSC Genome Browser Database (Karolchik et al. 2003). We then mapped these locations to the promoter sequences of the genes in our data set, based on genomic locations of transcriptional start sites, as reported in the GENCODE human alignment (Harrow et al. 2012). The vast majority (93%) of our 5 kbp-long promoter sequences contains DNase hypersensitive sites, and 60 percent of repeats overlap with a DNase hypersensitive site. This suggests that many of our repeat sequences could potentially be involved in gene regulation.

Genes with tandem repeats in promoters have significantly increased expression divergence.

We used previously published RNA-seq based gene expression data from our three study species (human, chimpanzee or macaque) and from different individuals (up to

six individuals per species), where each sample contained gene expression values for 13,035 genes in a given organ (brain, cerebellum, heart, kidney, liver or testis). We first asked if genes that contain tandem repeats in their promoters have higher expression divergence compared to genes without repeats in their promoter region. To this end, we computed the mean of gene expression values belonging to different individuals for each gene and organ. For each pair of species, we then calculated the difference between the mean expression values of the orthologous gene pairs, normalized by the sum of the mean expression values in a given organ. We then partitioned these pairwise expression differences into two subsets according to whether orthologous genes did or did not contain tandem repeats in their promoters. We observed a significant increase in pairwise expression differences of genes with repeats. More specifically, human-chimpanzee orthologs with repeats had higher mean expression difference ($P < 10^{-6}$, based on Wilcoxon Rank Sum Test (Mann and Whitney 1947)) compared to those without repeats. Similarly, human-macaque orthologs ($P < 0.01$, for all organs) and chimpanzee-macaque orthologs ($P < 10^{-5}$, for all organs) with repeats showed a higher mean expression difference than orthologous genes without repeats.

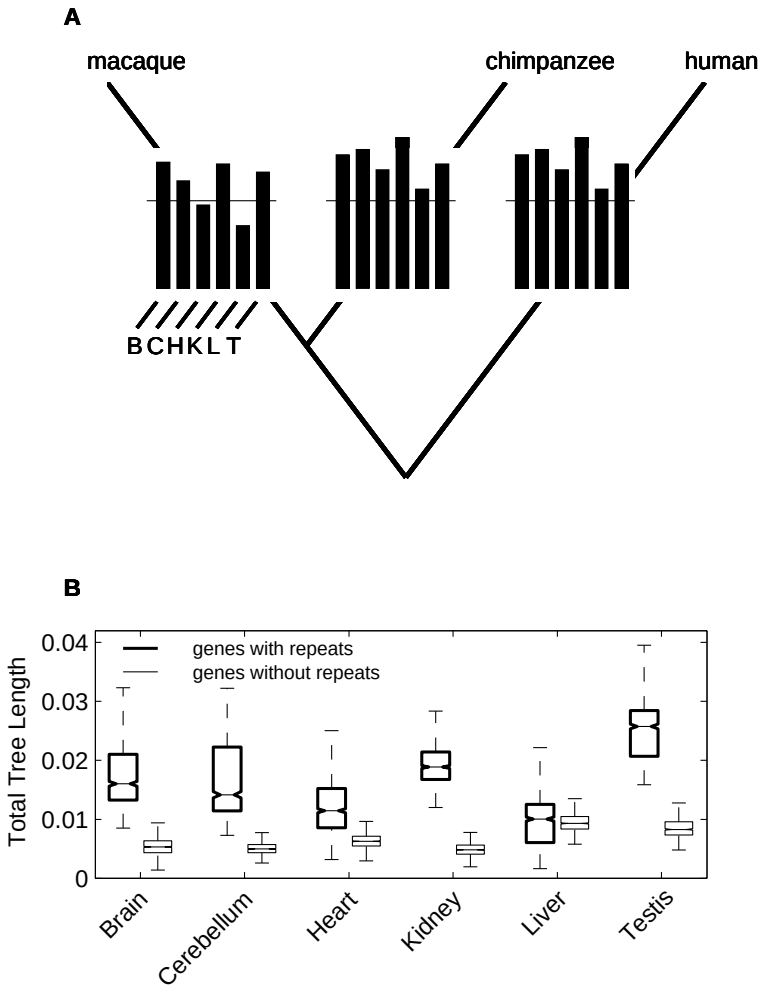

Figure 1. Presence of tandem repeats in promoters associate with increased expression divergence. (A) Schematic phylogenetic tree of our three study species with a bar-chart superimposed on each branch. The length of each bar indicates the ratios of branch lengths in 1000 sampled gene expression trees for genes with repeats relative to genes without repeats. Bars in each chart, from left to right, correspond to expression divergence in brain (B), cerebellum (C), heart (H), kidney (K), liver (L), and testis (T), respectively. Bars extending above the horizontal line indicate that genes with repeats show greater expression divergence. **(B)** Box plot of total tree lengths of genes with repeats (thick lines) and genes without repeats (thin lines). Horizontal lines in the middle of each box mark the median, edges of boxes correspond to the 25th and 75th percentiles, and whiskers cover 99.3 percent of the data points.

The preceding analysis, albeit simple and intuitive, can overestimate noise and underestimate organ-specific gene expression variation differences. In order to avoid these drawbacks, we next took a phylogenetic approach and performed a bootstrap-like resampling analysis, where gene expression values were sampled from different individuals of a species (see Material and methods). We used 1000 bootstrap-like replicates in this analysis, each with 13,035 sampled gene expression values. We then calculated the expression distance between each species pair separately for genes with repeats and without repeats for six different organs. Through this procedure, we arrived at 2 different expression distance matrices of (1000 replicates) × (3 species pairs) for each organ. We used these matrices to construct neighbor-joining gene expression trees. The constructed are thus unrooted trees, which have three branches that lead to the human, the chimpanzee, and the macaque lineage. The lengths of these branches indicate the amount of expression change that took place in each of the three lineages.

Figure 1A summarizes the ratios of branch lengths in these expression trees for genes with repeats relative to genes without repeats. Each bar in each chart corresponds to the branch lengths of an organ–specific gene expression tree for the lineage leading to this species. Specifically, from left to right, bars reflect gene expression divergence in brain (B), cerebellum (C), heart (H), kidney (K), liver (L), and testis (T). Bars extending above the horizontal line indicate that genes with repeats show greater expression divergence. Except for the macaque branch for liver- and heart-specific expression trees, all branches are significantly longer for repeat-containing genes ($P < 10^{-10}$; based on a t-test with $N = 1000$, df = N-1, throughout unless otherwise mentioned). Figure 1B indicates, separately for each organ, the distribution of total

tree length (summed over all three branches) for the 1000 bootstrapped trees. The total tree length of genes with repeats is significantly greater in all organs ($P < 10^{-200}$ except for liver, where $P = 0.02$). This analysis is based on repeats found in human genes. An analogous analysis with repeats found in the other two species shows that for most organs, genes with repeats have diverged to a significantly greater extent in those species as well. Specifically, $P < 10^{-207}$, in chimpanzee except for liver and testis, where *P*-values were non-significant (n.s.); and $P < 10^{-300}$ for macaques, except for testis (n.s.) (see supplementary figure S1, Supplementary Material online). Greater expression divergence in repeat containing genes persists also when we calculate expression distance with other methods (supplementary text S1, figure S2, Supplementary Material online) or when we correct for higher expression divergence in chimpanzee (supplementary text S2, figure S3, Supplementary Material online).

Most of our analysis relies on the program Tandem Repeat Finder (Gelfand et al. 2007). To confirm that our analysis is robust to alternative means of repeat detection, we used the algorithm implemented in Phobos (http://www.rub.de/spezzoo/cm/cm_phobos.htm) to detect human repeats. This analysis also showed that expression trees of genes with tandem repeats have significantly longer branches ($P < 10^{-9}$, except for testis (n.s.); see supplementary figure S4, Supplementary Material online).

Gene duplicates with tandem repeats have significantly increased expression divergence.

Gene duplication and subsequent divergence in gene expression is an important means of creating genes with new and specialized functions (Conant and Wolfe 2008;

Dong et al. 2011; Ganko et al. 2007; Gu et al. 2002; Hanada et al. 2009; Leach et al. 2007; Li et al. 2005). Because our previous analysis showed that the presence of tandem repeats increases expression divergence in general, we wondered whether repeats could also be associated with increased expression divergence in duplicate genes. To this end, we identified 8531 genes with one or more duplicates in our human gene data set (see Material and methods). Because some of these genes had up to 10 duplicates (mean number of duplicates: 2.9) the duplicates yielded a data set of 12,176 gene duplicate pairs. Of these, 48 percent (5879 pairs) had no repeats in their promoters, 42 percent (5130) had one or more tandem repeats in one copy, and in the remainder (1167) both duplicates had tandem repeats in their promoters.

We wanted to find out whether gene duplicates with repeats show increased expression divergence. To this end, we repeated our bootstrap-like analysis (see Material and methods) to generate 1000 replicates of our gene expression data, where the gene expression level of each human gene duplicate was sampled from different individuals among replicates. We then calculated expression distance matrices of size (1000×2) for our three categories of gene duplicates. For each of the six organs, we found that gene duplicates where both members carried repeats are much more divergent than gene duplicates where no member carried a repeat ($P < 10^{-261}$; see Figure 2). Gene pairs where only one member has a repeat showed greater expression divergence than gene pairs with no repeats ($P < 0.05$).

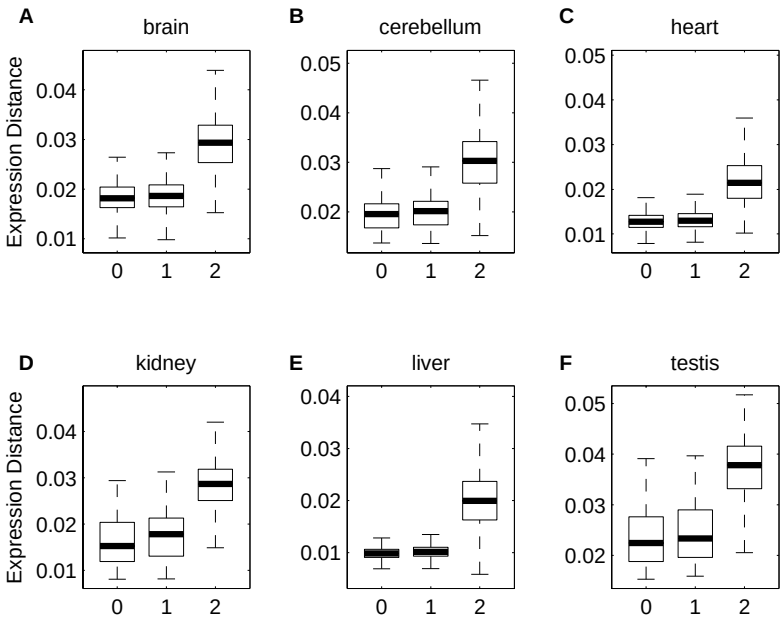

Figure 2. Presence of tandem repeats in duplicate genes associate with increased expression divergence. Box plot of total tree lengths of gene duplicates without repeats (left-most box in each panel), with repeats in the promoter of one duplicate (middle box), in the promoter of both duplicates (right-most box) for **(A)** brain, **(B)** cerebellum, **(C)** heart, **(D)** kidney, **(E)** liver, **(F)** testis. Horizontal lines in the middle of each box mark the median. The edges of the boxes correspond to the 25th and 75th percentiles. Whiskers cover 99.3 percent of the data points.

Repeat-containing genes are not under relaxed selection.

We next wondered whether the higher expression divergence of repeat-containing genes was simply due to relaxed selection that these genes experience, which can be detected through analysis of sequence divergence in their coding region. To investigate this possibility, we decided to use d_N/d_S, which is the ratio of the number of nonsynonymous substitutions per non-synonymous site to the number of synonymous substitutions per synonymous site, as an indicator of selective pressure acting on a protein-coding gene. We downloaded d_N and d_S values of the genes in our data set with Ensembl's Biomart tool (Kinsella et al. 2011), and calculated the ratio

d_N/d_S for duplicate genes and single copy genes, and subdivided genes in both categories into genes with and without tandem repeats in their promoter. We then asked whether repeat-containing genes show a higher d_N/d_S ratio, and thus evidence for relaxed selection, compared to genes without repeats. The answer is no, both for single-copy genes (P = 0.57; N = 7912, df= N-1) and for multi-copy genes (P = 0.45; N = 5123).

Association between tandem repeats and expression divergence is not dependent on expression level.

Because gene expression levels may play a role in expression divergence (Lehner 2008; Macneil and Walhout 2011; Pilpel 2011), we asked whether the association between tandem repeats and expression divergence varies with expression level. To this end, we distinguished between highly and lowly expressed genes, and chose the median expression level of all genes among all individuals in a species (and separately for each organ) as a threshold for high and low expression. We then followed our previously explained procedure to calculate 24 expression distance matrices of size (1000×3). We generated 1000 separate expression trees for genes with high and low expression, genes with and without repeats, and expression data from each organ, and calculated the total tree lengths for these trees. We then pooled tree lengths of different organs and asked whether total tree lengths differentiate to a similar extent between genes with repeats and genes without repeats for highly and lowly expressed genes. The answer is yes. Tandem repeats associate with expression divergence to a similar extent for highly expressed genes, ($P < 10^{-117}$, all organs considered together; N = 6000; df = N-1) as they do for lowly expressed genes ($P < 10^{-129}$, N = 6000). We then computed the pairwise difference in tree lengths between

pairs of trees derived from genes with repeats and genes without repeats for both of the sets, where two members in each pair of trees were obtained from expression data sampled from the same individuals. The mean differences in tree lengths are statistically indistinguishable between highly expressed genes (95% confidence intervals: 0.0034, 0.0042) and lowly expressed genes (95% CI: 0.0036, 0.0042).

The association still holds when repeats in CpG-islands are removed.

CpG islands play an important role in mammalian gene regulation (Saxonov et al. 2006) and may thus affect gene expression divergence. Because the G+C content (supplementary figure S5A, Supplementary Material online) of tandem repeats increases for repeats close to the transcription start site, we suspected that the association between repeats and increased expression divergence might stem from CpG islands. To find out, we first asked if repeats overlap with CpG islands. We retrieved experimentally identified CpG island locations from ENCODE (Material et al. 2004) and allocated these sites to the promoter sequences of the genes in our data set, based on genomic locations of transcription start sites, as reported in the GENCODE human alignment (Harrow et al. 2012) (see supplementary figure S5B, Supplementary Material online). We found that only 6% (215) of the repeats we identified overlap with a CpG island, a fraction that may be too small to influence all of the associations we observe. Indeed, when analyzing expression divergence while excluding repeats overlapping with CpG islands, we found that the presence of repeats is still strongly associated with expression divergence ($P < 10^{-178}$, for all organs except for liver, where $P < 10^{-6}$). To compare CpG island-associated repeats with other repeats more directly, we pooled tree lengths of different organs and computed the pairwise difference between replicates of genes with repeats and genes

without repeats. We found that mean differences in tree lengths are statistically indistinguishable, when we considered all repeats (95% CI: 0.0098, 0.0110) or only non-CpG repeats (95% CI: 0.0098, 0.0120). Hence, the presence of CpG island repeats is not likely to be a confounding factor in our analysis.

Repeat-associated divergence is even stronger for orthologous repeats.

Because we use expression values coming from multiple species, we wondered whether the association between repeats and expression divergence is stronger for *orthologous repeats*, i.e., for repeats with the same repeat unit that are present in both members of an orthologous gene pair. To this end, we first identified those repeats where regulatory regions of orthologous gene triplets in all three species contained a repeat with the same repeat unit (see Material and methods). However the number of such repeats was too small (45) for analysis. We therefore focused in the rest of this analysis on 718 orthologous repeats shared only between human and chimpanzee genes.

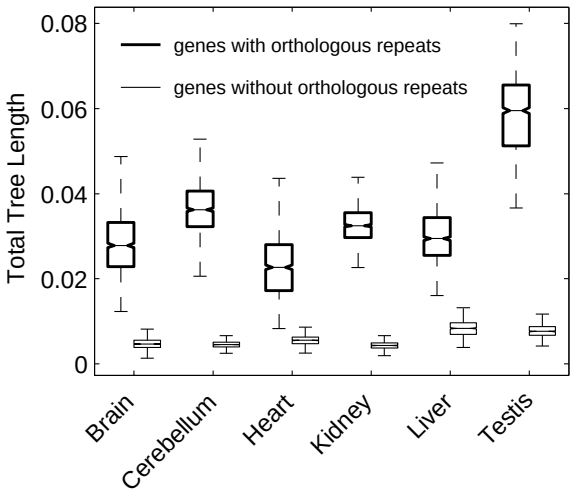

Figure 3. Orthologous repeats are associated more strongly with increased expression divergence. Box plot of total tree lengths of genes with repeats (thick boxes) and genes without orthologous repeats (thin boxes) for organ-specific gene expression, as indicated on the horizontal axis. Horizontal lines in the middle of each box mark the median. The edges of the boxes correspond to the 25th and

75th percentiles. Whiskers cover 99.3 percent of the data points.

After constructing expression trees for genes with orthologous repeats and genes without orthologous repeats, we calculated total tree lengths for the two gene sets. We found that genes with orthologous repeats have significantly higher expression divergence than genes without orthologous repeats ($P < 10^{-350}$ for each of the six organs; see Figure 3). We then asked whether the association is stronger for orthologous repeats than the association for all repeats (including the non-orthologous ones). To this end, we pooled tree length data from different organs and computed the pairwise difference in tree-length for genes with orthologous repeats and genes without orthologous repeats. The mean difference (95% CI: 0.0290, 0.0288) was significantly higher than the mean differences computed based on all repeats in human (95% CI: 0.0098, 0.0110) or in chimpanzee (95% CI: 0.0083, 0.0087).

Genes with the highest divergence are two fold more enriched with orthologous repeats than genes with lowest divergence.

We next asked whether genes with especially high expression divergence are especially highly enriched with tandem repeats. Our calculation of expression divergence so far had depended on tree lengths, which provides divergence information only for a set of genes, not for individual genes. For this analysis, it was necessary to distinguish individual genes with differential expression. We did so with the aid of the R package edgeR (Robinson et al. 2010), which estimates the extent of differential regulation of each gene, based on sharing information across the whole dataset using an empirical-Bayes-like procedure (Smyth 2004). To increase the quality of its divergence estimate, we used an expanded primate expression data set

(Brawand et al. 2011) that contains expression values not only from human, chimpanzee, and macaque, but also from gorilla and orangutan. We computed the extent of differential regulation for each gene in our data set and used the 1000 genes with the highest and lowest divergence in each organ for our analysis. In a majority of organs, genes with the highest divergence were twofold more enriched with orthologous repeats ($P < 0.01$ except for kidney and testis, where the enrichment was not significant.) When repeating our analysis with all repeats (including the non-orthologous repeats), we did not encounter a significant enrichment in any of the organs.

Repeats closer to the transcription start site show a stronger association with expression divergence.

Next we asked how our results would be affected if we changed the length of the upstream regions we consider. We thus identified human genes that contain repeats in upstream windows of length 1kbp (2404 genes with repeats), 10kbp (8736), 15kbp (10,056), and 20kbp (10,971; 2064 genes without repeats). As in our earlier analyses, we constructed expression trees based on repeat-containing and non-repeat-containing genes and compared the tree lengths for each window. These tree lengths were significantly different for windows of length 1 kbp ($P < 10^{-226}$), 10 kbp ($P < 10^{-12}$, except for liver, which shows an opposite trend with $P < 10^{-145}$), 15 kbp ($P < 10^{-15}$, except for liver, which shows an opposite trend with $P < 10^{-295}$) and 20 kbp (t-test; $P < 10^{-58}$, except for liver, which shows an opposite trend with $P < 10^{-350}$; see supplementary figure S6, Supplementary Material online). Figure 4 shows the mean difference in total expression tree length between repeat-containing and non-repeat-containing genes, based on 6000 gene expression trees when all organs are considered.

The difference is always positive, i.e., repeat-containing genes diverge more rapidly, but it is most pronounced for repeats 1kbp upstream of the transcription start site (95% CI: 0.0145, 0.0146). The difference gets progressively smaller as we include repeats that are further away from the transcription start site (95% CI for windows of length 10 kbp: 0.003, 0.006; 15 kbp: 0.0021, 0.0024; 20 kbp: 0.0004, 0.0007).

Repeats in other genic regions are also associated with increased expression divergence.

Although most transcriptional regulation is exerted by promoters, (Castillo-Davis et al. 2004; Ganapathi et al. 2007; Leach et al. 2007; Rockman and Wray 2002; Spitz and Furlong 2012; Wray et al. 2003a), sequences downstream of the gene, and especially 3' untranslated regions (3' UTRs) can also play an important role in gene regulation. We therefore wondered if repeat-containing 3'UTRs are also associated with higher expression divergence. To this end, we identified human genes in our data set that contained repeats within 1 kb of the 3' UTR There are 647 such genes, and they show significantly greater expression divergence in all organs except the testis ($P < 10^{-47}$, for all organs; supplementary figure S7A, Supplementary Material online).

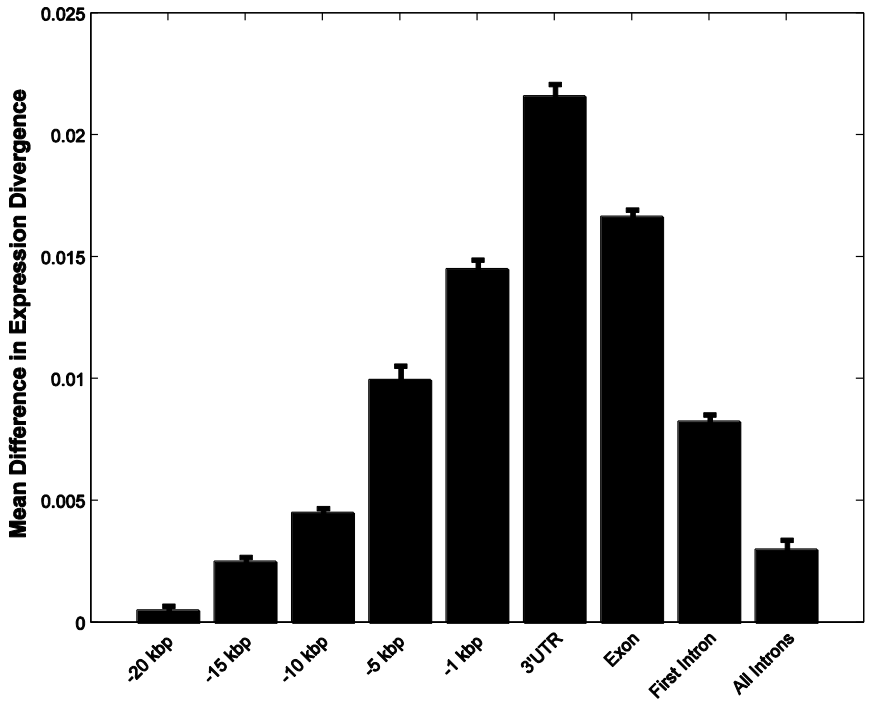

Figure 4. Presence of tandem repeats associate with higher expression divergence. Bars present mean differences in expression divergence, based on pairwise expression tree length differences between repeat-containing and non-repeat-containing genes. Repeats found in upstream regions of length 20kbp, 15kbp, 10kbp, 5kbp, 1 kbp, as well as in 3'UTRs, exons, first introns and all introns were considered, as indicated on the horizontal axis. Note that all expression differences are positive, indicating that repeat-containing genes, regardless of category, diverge more rapidly. Whiskers represent 95 percent confidence intervals.

Next we extended our analysis to repeats in exons and introns, because regulatory regions can also occur in these sequences (Charron et al. 2007; Gemayel et al. 2010; Jonsson et al. 1992; Rohrer and Conley 1998; Stranger et al. 2007a). We identified 2468 human genes with exon-containing repeats and found that they are associated with greater expression divergence in all organs ($P < 10^{-269}$, for all organs; supplementary figure S7B, Supplementary Material online). Furthermore, we analyzed 1336 human genes with repeats in their first intron, and 9521 genes with repeats in at least one intron, regardless of its location. We found that repeats in the

first intron are associated with high expression divergence ($P < 10^{-158}$) for all organs except for liver, which shows the opposite ($P < 10^{-3}$; supplementary figure S7C, Supplementary Material online). When we considered repeats found in all introns, repeat-containing genes showed again greater expression divergence ($P < 10^{-198}$) for all organs except for the heart ($P < 10^{-85}$) and the liver ($P < 10^{-292}$), both of which show opposite patterns (supplementary figure S7D, Supplementary Material online). However, the mean difference (95% CI: 0.0026, 0.0034) of tree lengths between repeat-containing and non-repeat-containing genes for repeats found in any intron was smaller compared to the mean difference for the repeats found in first introns (95% CI: 0.0080, 0.0086). Figure 4 illustrates that repeats found in 3'UTRs are associated with the strongest expression divergence (95% CI: 0.021, 0.022), followed by repeats found in exons (95% CI: 0.0166, 0.0167).

4.4. Discussion

Previous studies showed that tandem repeats found in various gene locations can change gene expression levels (Bennett et al. 1995; Fondon and Garner 2004; Hamada et al. 1984; Hammock and Young 2005; Lesch et al. 1996; Streelman and Kocher 2002). Here, we extended such analyses of individual genes to thousands of genes expressed in several organs of three primate model species, namely macaque, chimpanzee and human, and to gene expression divergence on evolutionary time scales. We found that the presence of repeats in gene promoters is strongly associated with evolutionary gene expression divergence, an observation that is robust to changes in the method to identify tandem repeats and to assess gene expression

divergence. The association exists for most of all organ-specific expression data, except for some analyses in testis and liver. Similar distinct expression patterns in these organs have been observed also by others (Brawand et al. 2011; Hsieh et al. 2003; Somel et al. 2008) in different contexts.

Repeats that are closer to the transcription start site are associated with greater expression divergence, an observation that can be explained through core promoter modules that occur preferentially close to this site and exert strong influence over transcriptional regulation (Spitz and Furlong 2012; Wray et al. 2003a). Moreover, an association with expression divergence holds also for repeats in other genic regions (Figure 4). The strongest of them is evident for 3'UTRs, consistent with their known role in gene regulation (Yoon et al. 2012). In addition, repeats in first introns are associated with greater expression divergence than repeats in other introns. This observation is consistent with previous work showing that most intronic regulatory regions occur in the first intron (Rohrer and Conley 1998), and that the first intron influences gene expression more than others (Charron et al. 2007; Jonsson et al. 1992). Taken together, our observations suggest an important role for tandem repeats in gene expression evolution.

Tandem repeats in the human genome are perhaps best-known for their pathological effects. Examples of diseases caused by variable tandem repeats are numerous and include Fragile X Syndrome, Huntington disease and spinobulbar muscular atrophy (Gemayel et al. 2010; Pearson et al. 2005a). Our work is consistent with previous analyses demonstrating that not all phenotypic variability that tandem repeats confer is deleterious, at least on an evolutionary time scale. A compelling example in

animals comes from the *Runx-2* gene, a major regulator of osteoblast differentiation. In human, repeat copy number variation in the promoter and exon of this gene cause cleidocranial dysplasia, a syndrome characterized by a variety of craniofacial and other skeletal malformations (Lee et al. 1997), whereas repeat variation in the dog *Runx-2* ortholog is a major source of non-pathological dug skull variation (Fondon and Garner 2004). Non-pathological expression variation associated with repeats has also been demonstrated in microbes. Specifically, (Vinces et al. 2009) showed that yeast genes with tandem repeats in their promoters have high expression divergence. And although a recent study (Elmore et al. 2012) in two different fungal species (*Aspergillus flavus* and *Apergillus oryzae*) found little evidence for such an association, the rarity of tandem repeats in fungal promoters (found in 2 percent of gene promoters, as opposed to some 30 percent in primate promoters) may be partly responsible.

Even though no study of associations can prove causation, we analyzed confounding factors that might have been at the root of the associations we observe. One of them was relaxed selection. Earlier work had detected that an increase in expression divergence for genes associated with species-specific transposable elements was caused by relaxed selection on those genes, rather than by the transposable elements themselves (Warnefors, Pereira, & Eyre-Walker, 2010). To exclude this factor, we showed that repeat-containing genes are not subject to relaxed purifying selection on their coding sequence. While we cannot exclude with certainty that relaxed selection acts only on the expression level of genes, we think this is unlikely, because we also analyzed the association between repeat presence and expression divergence in genes with high and low expression levels, and found no difference. A second possible

confounding factor is the presence of CpG islands, which are known to influence gene regulation, and which may cause the association we observed if many such islands overlap tandem repeats. However, we found out that this is not the case when we removed repeats containing CpG islands from our analysis, and showed that the association persists.

In a seminal paper, King and Wilson (King and Wilson 1975) observed about humans and chimpanzees that "their macromolecules are so alike that regulatory mutations may account for their biological differences." Since then, we have learned that such mutations, and in particular mutations that cause gene expression change, are indeed important in the evolution of primates and other organisms (Dimas et al. 2009; Fondon et al. 2008; Gemayel et al. 2010; Stranger et al. 2007a; Stranger et al. 2005; Vinces et al. 2009; Wren et al. 2000). Our work shows that tandem repeats and their high mutability may be an important class of regulatory mutations that are responsible for such species differences.

4.5. Methods

Gene expression and sequence data

The gene expression data we used is based on RNA sequencing of ~3.2 billion 76-base pair-long Illumina Genome Analyser IIx reads (Brawand et al. 2011). Expression levels are indicated as log2-transformed reads per kilobase of exon model per million mapped reads. It provides one-to-one gene expression measurements from multiple

primates, where each genes' expression had been measured in six different organs (brain, cerebellum, heart, kidney, liver, testis) and for 1-6 individuals per species, depending on species and organ (Brawand et al. 2011). From this data set, we used RNA-seq based expression values of all 13,035 one-to-one gene orthologs from humans, chimpanzees and macaques. We obtained DNA sequences of the genes in our expression data set through the Biomart tool of Ensemble (Kinsella et al. 2011), using human annotation version GRCh37.p10, chimpanzee annotation CHIMP2.1.4, and macaque annotation MMUL_1.0.

Tandem Repeat Identification

We identified tandem repeats in various regions of the genes we studied. These included the promoter (5,000 base pairs [bps] upstream from the transcription start site, unless stated otherwise), exons, the first intron, all introns, and the 3' untranslated region (1000 bps downstream from each gene's stop codon). We considered both micro- and minisatellites with tandem repeat units up to 50 nucleotides in length. Longer repeats are less variable and therefore less likely to cause phenotypic divergence (Kelkar et al. 2008; Li et al. 2004; Li et al. 2002; O'Dushlaine and Shields 2008; Payseur et al. 2011). We identified repeats with the program Tandem Repeat Finder v2.30 (Gelfand et al. 2007). Specifically, we analyzed only repeats with Tandem Repeat Finder scores exceeding 80, an incidence of indels in adjacent repeat units below 10 percent (e.g., a repeat unit of 20 nucleotides can have up to two indels relative to the consensus pattern, that is the repeat unit most similar to the whole repeat sequence (Gelfand et al. 2007)), and sequence identity of adjacent repeat units above 90 percent (e.g., at least 18 nucleotides of a repeat unit of 20 nucleotides must match the consensus pattern). One

motivation for these stringent thresholds is that the variability of tandem repeats increases strongly for repeats of high sequence similarity and Tandem Repeat Finder Scores (O'Dushlaine and Shields 2008). To validate the robustness of our results to the repeat identification process, we used another tool with a higher repeat detection power (Schaper et al. 2012), the Phobos 3.3.12 Tandem Repeat Search Tool (http://www.rub.de/spezzoo/cm/cm_phobos.htm) with the same match and indel criteria for repeat identification.

Identification of Orthologous Repeats

In the list of identified repeats based on the above criteria, we designated repeats that have the same repeat units and that lie upstream of orthologous genes as *orthologous repeats*. While it is in principle possible that the same orthologues could contain different repeats with the same repeat unit, there are only 10 such genes. This means that our rate of false positive orthologous repeat detection is very small. We allowed positional variation of repeats by up to 250 nucleotides, because promoter sequences and thus the position of regulatory elements can show substantial variation due to indels, even between closely related species (Hu and Ng 2012).

Calculation of Expression Divergence

The gene expression data set we used (Brawand et al. 2011) contains gene expression measurements from several individuals of a species for each gene and organ. We took advantage of this fact to assess statistical differences in gene expression divergence with a bootstrap-like resampling procedure (Brawand et al. 2011), where we sampled

gene expression values from different individuals of a species to create 1000 replicate data sets (n = 13,035) for each organ, and species.

We partitioned gene pairs in each such data set into two groups: gene pairs where genes of a given species contained tandem repeats in a specific region of interest, such as a promoter, and gene pairs without such repeats. We then computed, separately for genes in the two groups, a pairwise matrix of Euclidean gene expression distance between all genes in a pair of species, based on the formula (Tirosh et al. 2006)

$$ED_{i,j}(g) = \sqrt{\frac{1}{A_g}\sum_{k=1}^{A_g}\left(\frac{x_i(g,k)-x_j(g,k)}{x_i(g,k)+x_j(g,k)+2}\right)^2},$$

where i and j stand for species i and j (e.g. human and chimpanzee), g is a (binary) indicator variable reflecting which of two sets of genes (with or without repeats) are analyzed, A_g is the number of genes in that set, k is a gene-specific index and x is the expression level of a gene. To give an example, $x_{human}(no\ repeat, 1)$ is the gene expression value of the first gene in the human gene set without repeats for a given organ, and $ED_{\text{human, macaque}}$ (repeat) is the expression distance between repeat-containing human and macaque gene pairs for a given replicate data set.

Overall, we created 12 separate expression distance matrices of size (1000×3), for two gene subsets based on repeat presence and for six organs. We used these matrices to construct gene expression trees using the neighbor-joining approach (implemented in the 'ape' package (Paradis et al. 2004) in R (http://www.R-project.org/)). We used the branch lengths of the trees we constructed as a measure of gene expression divergence.

To test the null-hypothesis that the expression divergences (branch lengths) of the 1000 sampled trees were significantly different between the two gene subsets for each organ, we used paired t-tests (N=1000, df = N-1 unless otherwise mentioned). All P values are reported after Bonferroni correction (Dunn 1961) for multiple testing and they were robust to number of bootstrap replicates. We performed all statistical analyses using MATLAB (7.10.0, The MathWorks Inc., Natick, MA, R2010a).

Identification of Gene Duplicates

We obtained a list of gene duplicates in our human data set using the Biomart Tool in Ensemble (Kinsella et al. 2011). We excluded duplicates that were listed as merely "predicted paralogs" (as opposed to bona fide "within-species paralogs"), as well as duplicate pairs where the fraction K_a of non-synonymous substitutions per non-synonymous site exceeded one (Gu et al. 2004; Gu et al. 2003), because their divergence cannot be reliably estimated. We then grouped these duplicate pairs into three subsets, depending on whether none, one, or both of their member genes harbored repeats.

4.6. References

Bates, G. (1996). Expanded glutamines and neurodegeneration--a gain of insight. *BioEssays News and Reviews in Molecular Cellular and Developmental Biology*, *18*(3), 175–178. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=8867730

Bennett, S. T., Lucassen, A. M., Gough, S. C. L., Powell, E. E., Undlien, D. E., Pritchard, L. E., … Todd, J. A. (1995). Susceptibility to human type 1 diabetes at IDDM2 is determined by tandem repeat variation at the insulin gene minisatellite locus. *Nature Genetics*, *9*(3), 284–292.

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., … Kaessmann, H. (2011). The evolution of gene expression levels in mammalian organs. *Nature*, *478*(7369), 343–348. doi:10.1038/nature10532

Brinkmann, B., Klintschar, M., Neuhuber, F., Hühne, J., & Rolf, B. (1998). Mutation Rate in Human Microsatellites: Influence of the Structure and Length of the Tandem Repeat. *The American Journal of Human Genetics*, *62*(6), 1408–1415. Retrieved from http://dx.doi.org/10.1086/301869

Carroll, S. B. (2000). Endless forms: the evolution of gene regulation and morphological diversity. *Cell*, *101*, 577–580. doi:10.1016/S0092-8674(00)80868-5

Castillo-Davis, C. I., Hartl, D. L., & Achaz, G. (2004). cis-Regulatory and Protein Evolution in Orthologous and Duplicate Genes. *Genome Research*, *14*(8), 1530–1536. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=509261&tool=pmcentrez&rendertype=abstract

Charron, M., Chern, J.-Y., & Wright, W. W. (2007). The cathepsin L first intron stimulates gene expression in rat sertoli cells. *Biology of Reproduction*, *76*(5), 813–824. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17229931

Choi, J. K., & Kim, Y.-J. (2008). Epigenetic regulation and the variability of gene expression. *Nature Genetics*, *40*(2), 141–147. doi:10.1038/ng.2007.58

Conant, G. C., & Wolfe, K. H. (2008). Turning a hobby into a job: How duplicated genes find new functions. *Nature Reviews Genetics*, *9*(12), 938–950. doi:10.1038/nrg2482

Dimas, A. S., Deutsch, S., Stranger, B. E., Montgomery, S. B., Borel, C., Attar-Cohen, H., … Antonarakis, S. E. (2009). Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science*, *325*(5945), 1246–1250. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19644074

Dixon, A. L., Liang, L., Moffatt, M. F., Chen, W., Heath, S., Wong, K. C. C., … Cookson, W. O. C. (2007). A genome-wide association study of global gene expression. *Nature Genetics*, *39*(10), 1202–1207. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17873877

Dong, D., Yuan, Z., & Zhang, Z. (2011). Evidences for increased expression variation of duplicate genes in budding yeast: from cis- to trans-regulation effects. *Nucleic Acids Research*, *39*(3), 837–847. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3035465&tool=pmcentrez&rendertype=abstract

Dunn, O. J. (1961). Multiple Comparisons Among Means. *Journal of the American Statistical Association*, *56*, 52–64. doi:10.2307/2282330

Elmore, M. H., Gibbons, J. G., & Rokas, A. (2012). Assessing the genome-wide effect of promoter region tandem repeat natural variation on gene expression. *G3 (Bethesda, Md.)*, *2*(12), 1643–9. doi:10.1534/g3.112.004663

Fondon, J. W., & Garner, H. R. (2004). Molecular origins of rapid and continuous morphological evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(52), 18058–18063. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=539791&tool=pmcentrez&rendertype=abstract

Fondon, J. W., Hammock, E. a D., Hannan, A. J., & King, D. G. (2008). Simple sequence repeats: genetic modulators of brain function and behavior. *Trends in Neurosciences*, *31*(7), 328–34. doi:10.1016/j.tins.2008.03.006

Ganapathi, M., Singh, G. P., Sandhu, K. S., Brahmachari, S. K., & Brahmachari, V. (2007). A whole genome analysis of 5' regulatory regions of human genes for putative cis-acting modulators of nucleosome positioning. *Gene*, *391*(1-2), 242–251. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17321698

Ganko, E. W., Meyers, B. C., & Vision, T. J. (2007). Divergence in expression between duplicated genes in Arabidopsis. *Molecular Biology and Evolution*, *24*(10), 2298–309. doi:10.1093/molbev/msm158

Gelfand, Y., Rodriguez, A., & Benson, G. (2007). TRDB—The Tandem Repeats Database. *Nucleic Acids Research*, *35*(Database issue), D80–D87. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17175540

Gemayel, R., Vinces, M. D., Legendre, M., & Verstrepen, K. J. (2010). Variable Tandem Repeats Accelerate Evolution of Coding and Regulatory Sequences. *Annual Review of Genetics*. doi:10.1146/annurev-genet-072610-155046

Gu, Z., Nicolae, D., Lu, H. H.-S., & Li, W. H. (2002). Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends in Genetics : TIG*, *18*(12), 609–13. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12446139

Gu, Z., Rifkin, S. A., White, K. P., & Li, W.-H. (2004). Duplicate genes increase gene expression diversity within and between species. *Nature Genetics*, *36*(6), 577–9. doi:10.1038/ng1355

Gu, Z., Steinmetz, L., Gu, X., Scharfe, C., Davis, R., & Li, W. (2003). Role of duplicate genes in genetic robustness against null mutations. *Nature*, 63–66. doi:10.1038/nature01226.1.

Hamada, H., Seidman, M., Howard, B. H., & Gorman, C. M. (1984). Enhanced gene expression by the poly(dT-dG).poly(dC-dA) sequence. *Molecular and Cellular Biology*, *4*(12), 2622–2630. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=6098815

Hammock, E. a D., & Young, L. J. (2005). Microsatellite instability generates diversity in brain and sociobehavioral traits. *Science (New York, N.Y.)*, *308*(5728), 1630–4. doi:10.1126/science.1111427

Hanada, K., Kuromori, T., Myouga, F., Toyoda, T., & Shinozaki, K. (2009). Increased Expression and Protein Divergence in Duplicate Genes Is Associated with Morphological Diversification. *PLoS Genetics*, *5*(12), 7. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20041196

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., … Hubbard, T. J. (2012). GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Research*, *22*(9), 1760–1774. doi:10.1101/gr.135350.111

Hsieh, W.-P., Chu, T.-M., Wolfinger, R. D., & Gibson, G. (2003). Mixed-model reanalysis of primate data suggests tissue and species biases in oligonucleotide-based gene expression profiles. *Genetics*, *165*(2), 747–757. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1462792&tool=pmcentrez&rendertype=abstract

Hu, J., & Ng, P. C. (2012). Predicting the effects of frameshifting indels. *Genome Biology*. doi:10.1186/gb-2012-13-2-r9

Hurles, M. E., Dermitzakis, E. T., & Tyler-Smith, C. (2008). The functional impact of structural variation in humans. *Trends in Genetics : TIG*, *24*, 238–245. doi:10.1016/j.tig.2008.03.001

Jonsson, J. J., Foresman, M. D., Wilson, N., & McIvor, R. S. (1992). Intron requirement for expression of the human purine nucleoside phosphorylase gene. *Nucleic Acids Research*, *20*(12), 3191–3198. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=312458&tool=pmcentrez&rendertype=abstract

Jordan, I. K., Mariño-Ramírez, L., & Koonin, E. V. (2005). Evolutionary significance of gene expression divergence. *Gene*, *345*, 119–126. doi:10.1016/j.gene.2004.11.034

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T. S., Hinrichs, A., Lu, Y. T., … Kent, W. J. (2003). The UCSC Genome Browser Database. *Nucleic Acids Research*, *31*(1), 51–54. Retrieved from http://www.nar.oupjournals.org/cgi/doi/10.1093/nar/gkg129

Kashi, Y., & King, D. G. (2006). Simple sequence repeats as advantageous mutators in evolution. *Trends in Genetics*, *22*(5), 253–259. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16567018

Kelkar, Y. D., Tyekucheva, S., Chiaromonte, F., & Makova, K. D. (2008). The genome-wide determinants of human and chimpanzee microsatellite evolution, 30–38. doi:10.1101/gr.7113408.The

Kim, S.-J., Young, L. J., Gonen, D., Veenstra-VanderWeele, J., Courchesne, R., Courchesne, E., … Insel, T. R. (2002). Transmission disequilibrium testing of arginine vasopressin receptor 1A (AVPR1A) polymorphisms in autism. *Molecular Psychiatry*, *7*(5), 503–507. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12082568

King, M. C., & Wilson, A. C. (1975). Evolution at two levels in humans and chimpanzees. *Science*, *188*(4184), 107–116. doi:10.1126/science.1090005

Kinsella, R. J., Kähäri, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., … Flicek, P. (2011). Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database the Journal of Biological Databases and Curation*, *2011*(0), 9. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3170168&tool=pmcentrez&rendertype=abstract

Landry, C. R., Lemos, B., Rifkin, S. A., Dickinson, W. J., & Hartl, D. L. (2007). Genetic properties influencing the evolvability of gene expression. *Science*, *317*(5834), 118–121. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17525304

Leach, L. J., Zhang, Z., Lu, C., Kearsey, M. J., & Luo, Z. (2007). The role of cis-regulatory motifs and genetical control of expression in the divergence of yeast duplicate genes. *Molecular Biology and Evolution*, *24*(11), 2556–2565. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17846103

Lee, B., Thirunavukkarasu, K., Zhou, L., Pastore, L., Baldini, A., Hecht, J., … Karsenty, G. (1997). Missense mutations abolishing DNA binding of the osteoblast-specific transcription factor OSF2/CBFA1 in cleidocranial dysplasia. *Nature Genetics*, *16*, 307–310. doi:10.1038/ng0797-307

Legendre, M., Pochet, N., Pak, T., & Verstrepen, K. J. (2007). Sequence-based estimation of minisatellite and microsatellite repeat variability. *Genome Research*, *17*(12), 1787–1796. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2099588&tool=pmcentrez&rendertype=abstract

Lehner, B. (2008). Selection to minimise noise in living systems and its implications for the evolution of gene expression. *Molecular Systems Biology*, *4*(170), 170. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18319722

Lesch, K. P., Bengel, D., Heils, A., Sabol, S. Z., Greenberg, B. D., Petri, S., … Murphy, D. L. (1996). Association of Anxiety-Related Traits with a Polymorphism in the Serotonin Transporter Gene Regulatory Region. *Science*, *274*(5292), 1527–1531. doi:10.1126/science.274.5292.1527

Levinson, G., & Gutman, G. A. (1987). Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Molecular Biology and Evolution*, *4*(3), 203–221. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=3328815

Li, J., Liu, Y., Kim, T., Min, R., & Zhang, Z. (2010). Gene Expression Variability within and between Human Populations and Implications toward Disease Susceptibility. *PLoS Computational Biology*, *6*(8), 10. Retrieved from http://dx.plos.org/10.1371/journal.pcbi.1000910

Li, W.-H., Yang, J., & Gu, X. (2005). Expression divergence between duplicate genes. *Trends in Genetics*, *21*(11), 602–607. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16140417

Li, Y.-C., Korol, A. B., Fahima, T., Beiles, A., & Nevo, E. (2002). Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Molecular Ecology*, *11*(12), 2453–65. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12453231

Li, Y.-C., Korol, A. B., Fahima, T., & Nevo, E. (2004). Microsatellites within genes: structure, function, and evolution. *Molecular Biology and Evolution*, *21*(6), 991–1007. doi:10.1093/molbev/msh073

Macneil, L. T., & Walhout, A. J. M. (2011). Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. *Genome Research*, *21*(5), 645–57. doi:10.1101/gr.097378.109

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, *18*(1), 50–60. doi:10.1214/aoms/1177730491

Material, S. O., Web, S., Press, H., York, N., & Nw, A. (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science (New York, N.Y.)*, *306*(5696), 636–40. doi:10.1126/science.1105136

O'Dushlaine, C. T., & Shields, D. C. (2008). Marked variation in predicted and observed variability of tandem repeat loci across the human genome. *BMC Genomics*, *9*, 175. doi:10.1186/1471-2164-9-175

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, *20*(2), 289–290. Retrieved from http://www.bioinformatics.oupjournals.org/cgi/doi/10.1093/bioinformatics/btg412

Payseur, B. a, Jing, P., & Haasl, R. J. (2011). A genomic portrait of human microsatellite variation. *Molecular Biology and Evolution*, *28*(1), 303–12. doi:10.1093/molbev/msq198

Pearson, C. E., Nichol Edamura, K., & Cleary, J. D. (2005). Repeat instability: mechanisms of dynamic mutations. *Nature Reviews Genetics*, *6*(10), 729–742. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16205713

Pilpel, Y. (2011). Noise in biological systems: pros, cons, and mechanisms of control. *Methods In Molecular Biology Clifton Nj*, *759*, 407–425. Retrieved from http://www.springerlink.com/index/10.1007/978-1-61779-173-4

Ponting, C. P. (2008). The functional repertoires of metazoan genomes. *Nature Reviews Genetics*, *9*, 689–698.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2796818&tool=pmcentrez&rendertype=abstract

Rockman, M. V, & Wray, G. a. (2002). Abundant raw material for cis-regulatory evolution in humans. *Molecular Biology and Evolution*, *19*(11), 1991–2004. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12411608

Rohrer, J., & Conley, M. E. (1998). Transcriptional regulatory elements within the first intron of Bruton's tyrosine kinase. *Blood*, *91*(1), 214–221.

Saxonov, S., Berg, P., & Brutlag, D. L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(5), 1412–1417. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1345710&tool=pmcentrez&rendertype=abstract

Schaper, E., Kajava, A. V, Hauser, A., & Anisimova, M. (2012). Repeat or not repeat?--Statistical validation of tandem repeat prediction in genomic sequences. *Nucleic Acids Research*, *40*(20), 10005–17. doi:10.1093/nar/gks726

Schlötterer, C. (2000). Evolutionary dynamics of microsatellite DNA. *Chromosoma*, *109*(6), 365–371. doi:10.1007/s004120000089

Schlotterer, C., & Tautz, D. (1992). Slippage synthesis of simple sequence DNA. *NuclAcids Res*, *20*(2), 211–215. Retrieved from http://research.bmn.com/medline/search/record?uid=MDLN.92158603

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, *3*, Article3. doi:10.2202/1544-6115.1027

Somel, M., Creely, H., Franz, H., Mueller, U., Lachmann, M., Khaitovich, P., & Pääbo, S. (2008). Human and Chimpanzee Gene Expression Differences Replicated in Mice Fed Different Diets. *PLoS ONE*, *3*(1), 7. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18231591

Spitz, F., & Furlong, E. E. M. (2012). Transcription factors: from enhancer binding to developmental control. *Nature Reviews. Genetics*, *13*(9), 613–26. doi:10.1038/nrg3207

Stranger, B. E., Forrest, M. S., Clark, A. G., Minichiello, M. J., Deutsch, S., Lyle, R., … Dermitzakis, E. T. (2005). Genome-Wide Associations of Gene Expression Variation in Humans. *PLoS Genetics*, *1*(6), 10. Retrieved from http://discovery.ucl.ac.uk/1316078/

Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazley, C., Thorne, N., … Dermitzakis, E. T. (2007). Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, *315*(5813), 848–853. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2665772&tool=pmcentrez&rendertype=abstract

Stranger, B. E., Nica, A. C., Forrest, M. S., Dimas, A., Bird, C. P., Beazley, C., … Dermitzakis, E. T. (2007). Population genomics of human gene expression. *Nature Genetics*, *39*(10), 1217–1224. doi:10.1038/ng2142

Streelman, J. T., & Kocher, T. D. (2002). Microsatellite variation associated with prolactin expression and growth of salt-challenged tilapia. *Physiological Genomics*, *9*(1), 1–4. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=11948285

Tirosh, I., Barkai, N., & Verstrepen, K. J. (2009). Promoter architecture and the evolvability of gene expression. *Journal of Biology*, *8*(11), 95. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2804285&tool=pmcentrez&rendertype=abstract

Tirosh, I., Weinberger, A., Carmi, M., & Barkai, N. (2006). A genetic signature of interspecies variations in gene expression. *Nature Genetics*, *38*(7), 830–834. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16783381

Van Ham, S. M., van Alphen, L., Mooi, F. R., & van Putten, J. P. (1993). Phase variation of H. influenzae fimbriae: transcriptional control of two divergent genes through a variable combined promoter region. *Cell*, *73*, 1187–1196. doi:10.1016/0092-8674(93)90647-9

Vinces, M. D., Legendre, M., Caldara, M., Hagihara, M., & Verstrepen, K. J. (2009). Unstable tandem repeats in promoters confer transcriptional evolvability. *Science (New York, N.Y.)*, *324*(5931), 1213–6. doi:10.1126/science.1170097

Waal, F. B. M. De. (2009). *The Age of Empathy*. *Harmony Retrieved from httpwww goodreads combookshow6525532theageofempathy* (p. 291). Harmony Books. Retrieved from http://books.google.com/books?id=hmknDAdHYyEC

Weber, J. L., & Wong, C. (1993). Mutation of human short tandem repeats. *Human Molecular Genetics*, *2*(8), 1123–8. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8401493

Webster, M. T., Smith, N. G. C., & Ellegren, H. (2002). Microsatellite evolution inferred from human–chimpanzee genomic sequence alignments. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(13), 8748–8753. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=124370&tool=pmcentrez&rendertype=abstract

Wray, G. a, Hahn, M. W., Abouheif, E., Balhoff, J. P., Pizer, M., Rockman, M. V, & Romano, L. a. (2003). The evolution of transcriptional regulation in eukaryotes. *Molecular Biology and Evolution*, *20*(9), 1377–419. doi:10.1093/molbev/msg140

Wren, J. D., Forgacs, E., Fondon III, J. W., Pertsemlidis, A., Cheng, S. Y., Gallardo, T., … Garner, H. R. (2000). Repeat Polymorphisms within Gene Regions: Phenotypic and Evolutionary Implications. *The American Journal of Human Genetics*, *67*(2), 345–356. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1287183&tool=pmcentrez&rendertype=abstract

Yoon, O. K., Hsu, T. Y., Im, J. H., & Brem, R. B. (2012). Genetics and regulatory impact of alternative polyadenylation in human B-lymphoblastoid cells. *PLoS Genetics*, *8*(8), e1002882. doi:10.1371/journal.pgen.1002882

4.7. Supplementary Material

Text S1: The association between tandem repeat presence and expression divergence persists for different measures of expression distance.

In most of our analyses we used Tirosh's Euclidean distance (Tirosh et al. 2006) as a measure of gene expression divergence, because it does not amplify gene expression noise (Glazko and Mushegian 2010). Other ways to estimate gene expression distance include Pearson's correlation coefficient (Brawand et al. 2011; Meisel et al. 2012), various test statistics, such as that of the Kolmogorov-Smirnov test (Choi and Kim 2008), the t-test (Elmore et al. 2012; Li et al. 2010), and the Wilcoxon-Mann-Whitney test (Gu et al. 2004), as well as logarithmically transformed expression ratios (Enard et al. 2002). To determine how robust our observations are to changes in the divergence measure, we repeated our analysis with several other distance measures, including the Spearman's rank correlation coefficient r, as well as t-statistics and log-transformed ratios of expression values, as proposed by (Enard et al. 2002).

Results of these analyses (Supplementary Figure 4) suggest that the greater expression divergence of repeat-containing genes does not depend on the specific distance measure used. For example, a t-test of the null hypothesis that expression divergence of genes with repeats and genes without repeats are statistically indistinguishable is rejected at $P < 10^{-85}$ for a correlation-based distance measure, at $P < 10^{-350}$ for a t-statistic-based distance measure, and at $P < 10^{-350}$ for expression ratios, when the tree lengths of different organs are pooled in one matrix. Moreover, genes with repeats were significantly more diverged when we analyzed expression in different organs separately with a t-statistic-based expression distance ($P < 10^{-100}$) except for heart (n.s.) and liver (an opposite trend of $P < 10^{-5}$), with a correlation-based

approach ($P < 10^{-3}$, for each organ) and with an expression ratio approach ($P < 10^{-233}$, for each organ).

Text S2: The association between tandem repeats and expression divergence holds after correction for higher within-species expression divergence in chimpanzees.

Because chimpanzees may have greater within-species expression divergence than humans (e.g., Warnefors & Eyre-Walker (2012) reported a 2.5-fold difference) our analysis thus far may have overestimated the overall extent of expression divergence, especially for those genes with already high expression divergence. We therefore decided to repeat our analysis while correcting for potential differences in within-species divergences. As a crude estimate independent of that by Warnefors & Eyre-Walker (2012) for the expected ratio of gene expression divergence between human and chimpanzee, we used the ratio $2N_e/t$ (Hsieh et al. 2003), where N_e is the effective population size and t is the time elapsed (in generations) since their split. Based on effective population sizes of N_e = 10,000 (Hill 1981) and N_e = 25,000 for humans and chimps, respectively, (Eyre-Walker et al. 2002; Won and Hey 2005), as well as on generation times of 30 and 25 years for humans and chimps, respectively, (Langergraber et al. 2012), one would expect a 2.1 fold greater expression divergence for chimpanzees, which is similar to the previously observed value of 2.5 (Warnefors and Eyre-Walker 2012). Motivated by this analysis, we divided all chimpanzee expression values by 2.5, recalculated the expression distances for all species pairs as previously described and recomputed the expression trees. Comparison of total tree lengths between genes with repeats and genes without repeats suggest that tandem repeats are still associated with higher expression divergence ($P < 10^{-350}$, except for liver (n.s.); see Supplementary Figure 3).

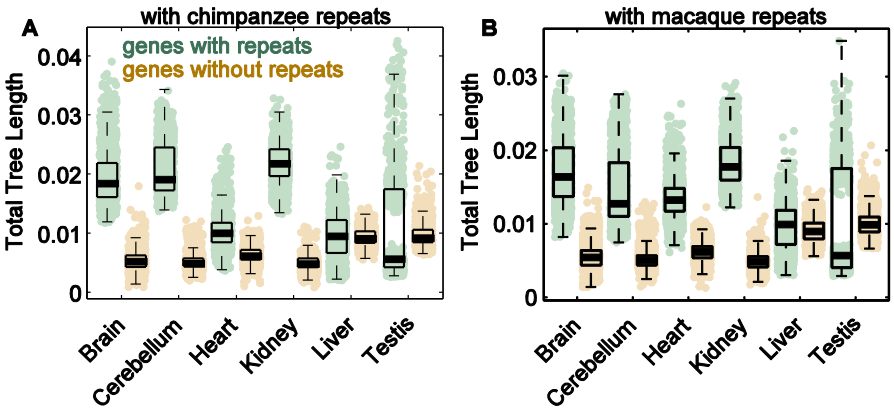

Figure S1. Expression divergence calculated through non-human repeats. Box plot of total tree lengths (vertical axes) of genes with repeats (green) and genes without repeats (orange) constructed based on **(A)** chimpanzee and **(B)** macaque repeats for organ-specific gene expression trees, as indicated on the horizontal axis. Each colored dot represents the length of one among the 1000 replicate gene expression divergence trees. The horizontal line in the middle of each box marks the median. The edges of each box correspond to the 25th and 75th percentiles. Whiskers cover 99.3 percent of the data points.

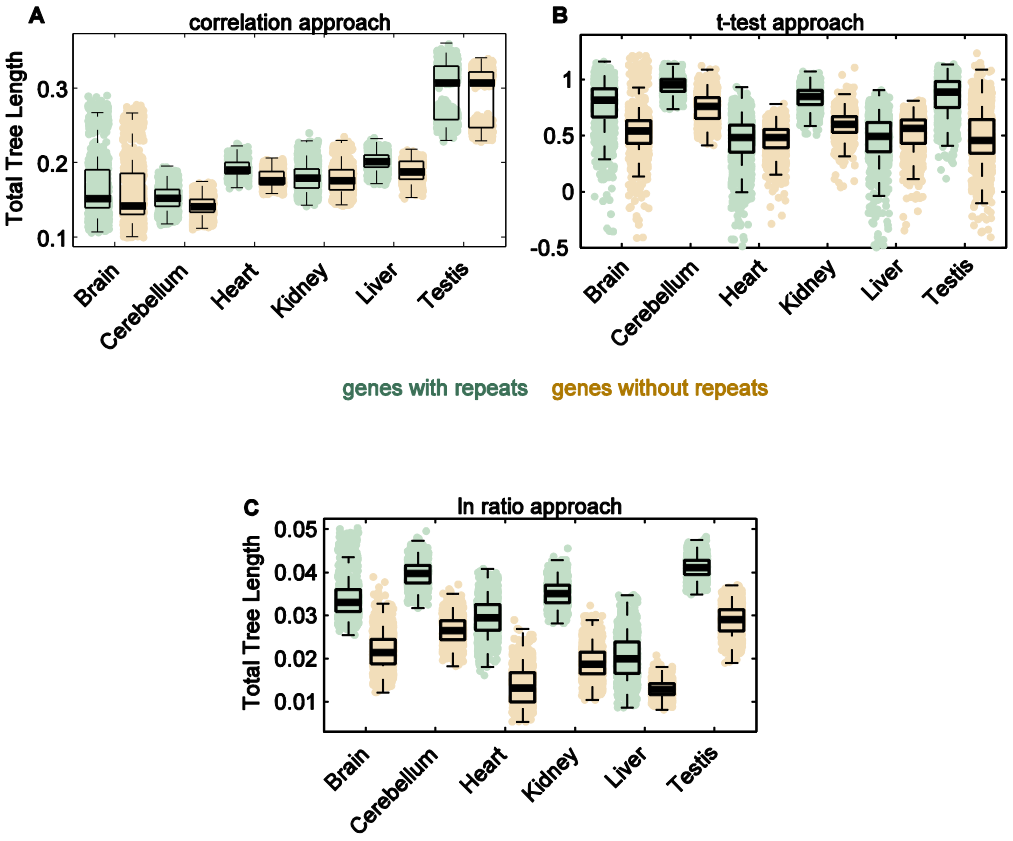

Figure S2. Expression divergence calculated by other approaches. The horizontal axes show the organs for which we constructed organ-specific gene expression trees. The vertical axes show the distribution of total gene expression divergence tree lengths of genes with repeats (green) and genes without repeats (orange), constructed by gene expression distances calculated from **(A)** correlation (Spearman's Rho) **(B)** t-test statistics **(C)** logarithm of expression ratios (as in (Enard et al. 2002)). Horizontal lines in the middle of each box mark the median. The edges of the boxes correspond to the 25th and 75th percentiles. Colored dots present the distribution of tree lengths for each of 1000 replicates. Whiskers cover 99.3 percent of the data points.

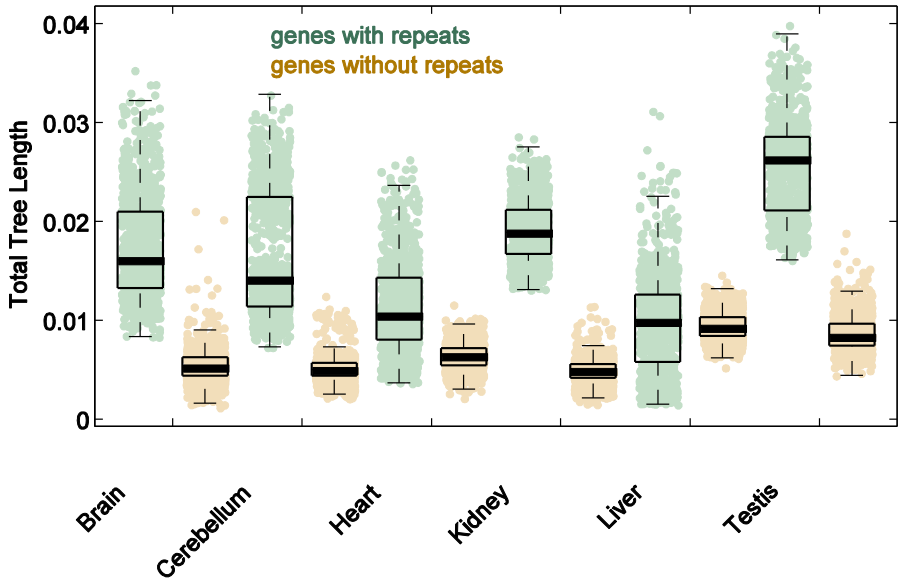

Figure S3. Expression divergence analysis after correction for higher divergence in chimpanzees. The horizontal axis shows the organs for which we constructed organ-specific gene expression trees. The vertical axis shows the distribution of total gene expression divergence tree lengths of genes with repeats (green) and genes without repeats (orange). Horizontal lines in the middle of each box mark the median. The edges of the boxes correspond to the 25th and 75th percentiles. Colored dots present the distribution of tree lengths for each of 1000 replicates. Whiskers cover 99.3 percent of the data points.

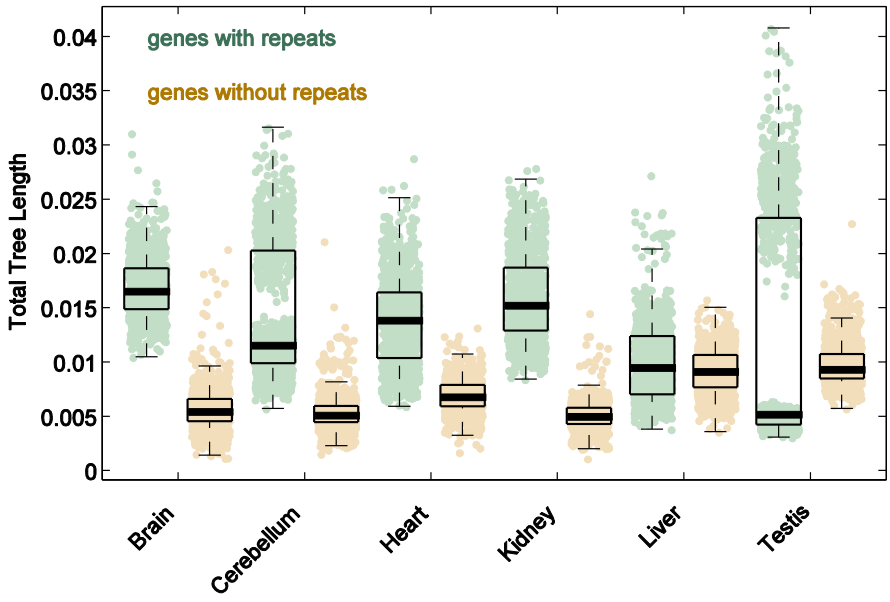

Figure S4. Expression divergence analysis based on repeats identified by Phobos. The horizontal axis shows the organs for which we constructed organ-specific gene expression trees. The vertical axis shows the distribution of total gene expression divergence tree lengths for genes with repeats (green) and genes without repeats (orange), as identified by Phobos. Horizontal lines in the middle of each box mark the median. The edges of the boxes correspond to the 25th and 75th percentiles. Pair of different colored boxes corresponds to the tree lengths of gene expression trees for their below specified organ. Colored dots present the distribution of tree lengths for each 1000 replicate. Whiskers cover 99.3 percent of the data points.

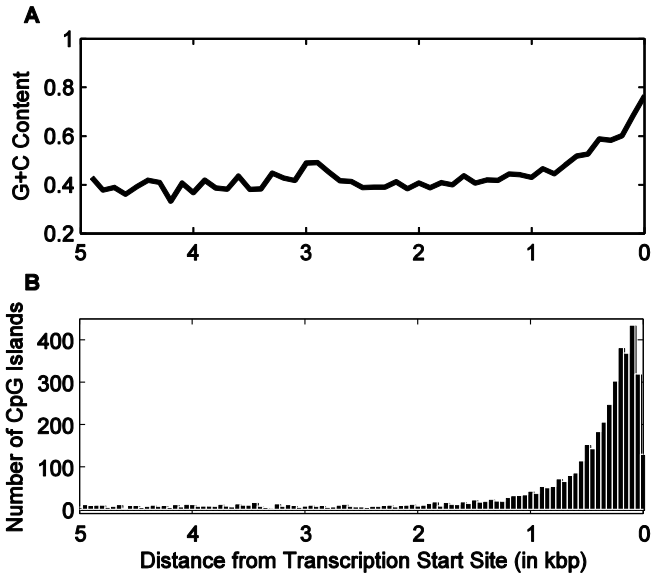

Figure S5. Higher G+C content and more CpG sites close to transcription start site. Plot of (A) (G+C) content of repeats, calculated as (G+C)/(C+G+A+T) based on the median of sliding windows that are 100 nucleotides long (B) number of CpG islands frequency (locations retrieved from ENCODE) in upstream regions of human genes.

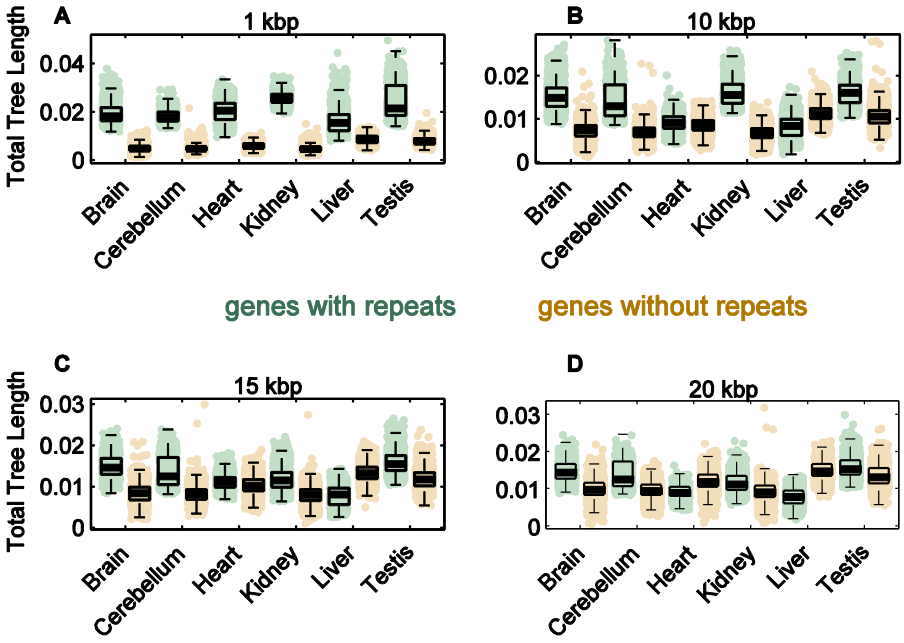

Figure S6. Repeats closer to transcription start site are associated more strongly with expression divergence. The horizontal axes show the organs for which we constructed organ-specific gene expression trees. The vertical axes show the distribution of total gene expression divergence tree lengths for genes with repeats (green) and without repeats (orange), where repeats could occur in upstream regions of length **(A)** 1kbp **(B)** 10kbp **(C)** 15kbp **(D)** 20kbp. Horizontal lines in the middle of each box mark the median. The edges of the boxes correspond to the 25th and 75th percentiles. Colored dots present the distribution of tree lengths for each of 1000 replicates. Whiskers cover 99.3 percent of the data points.

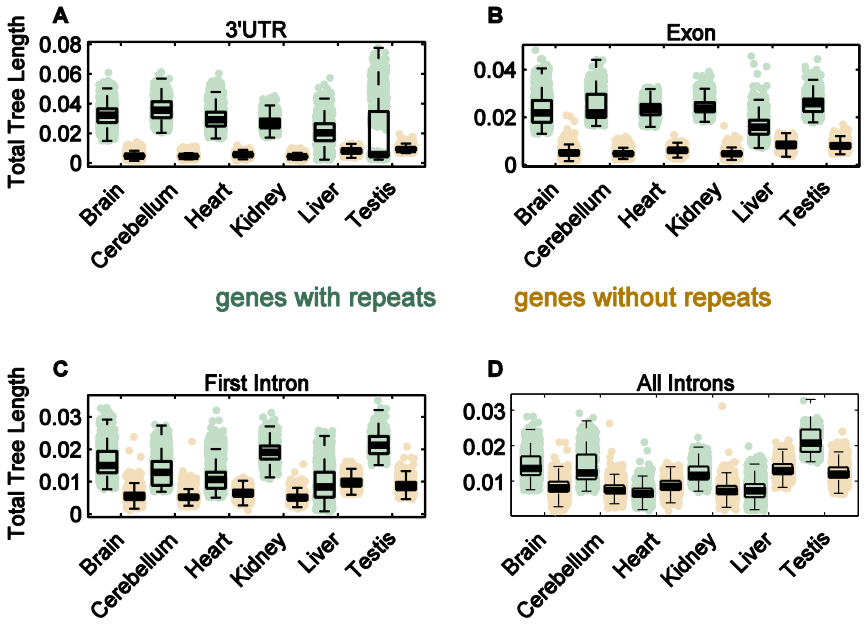

Figure S7. Tandem repeats are associated with increased expression divergence. The horizontal axes show the organs for which we constructed organ-specific gene expression trees. The vertical axes show the distribution of total gene expression divergence tree lengths for genes with repeats (green) and genes without repeats (orange), based on repeats found in **(A)** exons, **(B)** 3'UTR regions, **(C)** the first intron of a gene, **(D)** all introns. Horizontal lines in the middle of each box mark the median. The edges of the boxes correspond to the 25th and 75th percentiles. Colored dots present the distribution of tree lengths for each of 1000 replicates. Whiskers cover 99.3 percent of the data points.

References

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., … Kaessmann, H. (2011). The evolution of gene expression levels in mammalian organs. *Nature*, *478*(7369), 343–348. doi:10.1038/nature10532

Choi, J. K., & Kim, Y.-J. (2008). Epigenetic regulation and the variability of gene expression. *Nature Genetics*, *40*(2), 141–147. doi:10.1038/ng.2007.58

Elmore, M. H., Gibbons, J. G., & Rokas, A. (2012). Assessing the genome-wide effect of promoter region tandem repeat natural variation on gene expression. *G3 (Bethesda, Md.)*, *2*(12), 1643–9. doi:10.1534/g3.112.004663

Enard, W., Khaitovich, P., Klose, J., Zöllner, S., Heissig, F., Giavalisco, P., … Pääbo, S. (2002). Intra- and interspecific variation in primate gene expression patterns. *Science (New York, N.Y.)*, *296*(5566), 340–3. doi:10.1126/science.1068996

Eyre-Walker, A., Keightley, P. D., Smith, N. G. C., & Gaffney, D. (2002). Quantifying the slightly deleterious mutation model of molecular evolution. *Molecular Biology and Evolution*, *19*(12), 2142–2149. Retrieved from http://sro.sussex.ac.uk/20246/

Glazko, G., & Mushegian, A. (2010). Measuring gene expression divergence: the distance to keep. *Biology Direct*, *5*(1), 51. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2928186&tool=pmcentrez&rendertype=abstract

Gu, Z., Rifkin, S. A., White, K. P., & Li, W.-H. (2004). Duplicate genes increase gene expression diversity within and between species. *Nature Genetics*, *36*(6), 577–9. doi:10.1038/ng1355

Hill, W. G. (1981). Estimation of effective population size from data on linkage disequilibrium. *Genetical Research*, *38*(3), 209–216. Retrieved from http://journals.cambridge.org/abstract_S0016672300020553

Hsieh, W.-P., Chu, T.-M., Wolfinger, R. D., & Gibson, G. (2003). Mixed-model reanalysis of primate data suggests tissue and species biases in oligonucleotide-based gene expression profiles. *Genetics*, *165*(2), 747–757. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1462792&tool=pmcentrez&rendertype=abstract

Langergraber, K. E., Prüfer, K., Rowney, C., Boesch, C., Crockford, C., Fawcett, K., … Vigilant, L. (2012). Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *Proc Natl Acad Sci USA*, *In Press*. Retrieved from http://www.pnas.org/content/early/2012/08/08/1211740109.abstract

Li, J., Liu, Y., Kim, T., Min, R., & Zhang, Z. (2010). Gene Expression Variability within and between Human Populations and Implications toward Disease Susceptibility. *PLoS Computational Biology*, *6*(8), 10. Retrieved from http://dx.plos.org/10.1371/journal.pcbi.1000910

Meisel, R. P., Malone, J. H., & Clark, A. G. (2012). Faster-x evolution of gene expression in Drosophila. *PLoS Genetics*, *8*(10), e1003013. doi:10.1371/journal.pgen.1003013

Tirosh, I., Weinberger, A., Carmi, M., & Barkai, N. (2006). A genetic signature of interspecies variations in gene expression. *Nature Genetics*, *38*(7), 830–834. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16783381

Warnefors, M., & Eyre-Walker, A. (2012). A selection index for gene expression evolution and its application to the divergence between humans and chimpanzees. *PloS One*, *7*(4), e34935. doi:10.1371/journal.pone.0034935

Won, Y.-J., & Hey, J. (2005). Divergence population genetics of chimpanzees. *Molecular Biology and Evolution*, *22*(2), 297–307. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15483319

5. A survey of tandem repeat instability and gene expression changes in 37 colorectal cancers

Tugce Bilgin [1,2], Andreas Wagner [1,2,3]

[1]Institute of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland, [2] The Swiss Institute of Bioinformatics, Lausanne, Switzerland, [3] The Santa Fe Institute, Santa Fe, New Mexico, United States of America

5.1. Abstract

Colorectal cancer is a major contributor to cancer morbidity and mortality. A better understanding of tumor-associated molecular alterations is needed to gain insight into its carcinogenesis. Tandem repeat variation, a hallmark of colorectal cancer, and its effect on cancer phenotype remain so far poorly studied on a genome-wide scale. Here we present a systematical analysis of tandem repeat instability in the genomes of 37 colorectal tumors and their matched normal tissues for the upstream regulatory regions of 18,709 genes. We find that 5 percent of tandem repeats vary in copy number between tumor and their matched normal genomes, whereas between normal genome pairs only 2.7 percent of repeats vary. Furthermore, tumor/normal genome pairs show almost twice as much repeat loss or gain as normal genome pairs. We find that genes with repeat instability are significantly overexpressed. When we analyze well-studied cancer-associated signaling pathways, we find that most pathways are significantly enriched for repeat instability in tumor/normal pairs compared to normal genome pairs, and genes in these pathways with such unstable repeats are consistently overexpressed. Our results suggest an important role for promoter tandem repeat instability in differential gene expression of colorectal tumors.

5.2. Introduction

Microsatellites, short tandem DNA repeats, are among the most variable loci in the human genome, experiencing mutations in the copy number of repeat units at a rate of 10^{-3} to 10^{-7} per cell division (Legendre et al. 2007; Li et al. 2002). Most mutations giving rise to such unstable tandem repeats result from replication slippage that escaped the proofreading activity of mismatch repair systems (Schlötterer 2000). Repeat instability is associated with disease susceptibility and pathogenesis (Gemayel et al. 2010; López Castel et al. 2010; Vilar and Gruber 2010). It is common in many cancers, including colorectal, gastric, endometrial, ovarian, and breast cancer (Imai and Yamamoto 2008; Woerner et al. 2003). For example, a CAG tri-nucleotide repeat associated with prostate cancer has been identified in the first exon of the *androgen receptor* gene. Expansion of this repeat decreases gene expression, and increases disease incidence and tumor aggression (Giovannucci et al. 1997). In breast cancer, a dinucleotide CA-repeat within the first intron of the *epidermal growth factor receptor* (*EGFR*) gene correlates with the gene's transcription levels. Mutant alleles of the highly polymorphic 28 base pair long repeat in the downstream region of the proto-oncogene *HRAS1* significantly increases disease susceptibility for many cancers, including breast, colon, rectum, urinary, bladder cancer, and leukemia (Krontiris et al. 1993).

Colorectal cancer is the third most commonly diagnosed cancer in the world, and the second leading cause of cancer-related deaths in western societies (Siegel et al. 2013; UK 2014). Despite a large number of studies on colorectal cancer treatment, current therapeutic approaches cure only a fraction of patients (Hewish et al. 2010; Vilar and

Gruber 2010), which necessitates a better understanding of molecular alterations causing carcinogenesis. Colorectal cancers are initiated by several genetic alterations, including inactivation of p53, a tumor-suppressor gene mutated in most cancer types (Kan et al. 2010) and of the *adenomatous polyposis coli* (*APC*) gene, a key member of the Wnt signaling pathway (Van Limbergen et al. 2002). Mutations in *APC* silence the gene's expression, which leads to uncontrolled cell growth (Van Limbergen et al. 2002). Another gene in the Wnt pathway, *Dickkopf-3* (*DKK3*) helps microvessels to feed cancer cells and is expressed only in angiogenic tumors (Zitt et al. 2008). Like the Wnt pathway, mTOR pathway, which regulates cell growth and survival is activated in most colorectal cancers (Laplante and Sabatini 2012; Wang and Zhang 2014). Increased expression levels of *mTOR* correlate well with cancer stage (Alqurashi et al. 2013).

Several gene expression profiling studies of colorectal adenomas showed that tumors with mutations in different genes have distinctive expression patterns (Di Pietro et al. 2005; Tian et al. 2012). The patterns detected from such large-scale gene expression data sets are already being used to stratify tumor subtypes and to predict patient survival (Burgess 2013; Nosho et al. 2005). A study on comparability of gene expression changes in colorectal cancer, based on data produced in different laboratories showed that on average 95 percent of genes show consistent gene expression changes between two major subtypes of colorectal cancer, independent of the source of the data (Jorissen et al. 2008). The importance of gene expression patterns in tumor characterization requires a more complete understanding of the kinds of mutations that contribute to tumor-specific gene expression divergence.

Colorectal cancers are associated with chromosomal instability and microsatellite instability, which are not mutually exclusive (Imai and Yamamoto 2008). Microsatellite instability is found in at least 15 percent of sporadic colorectal cancers and is the major characteristic of hereditary colorectal cancer (Vilar and Gruber 2010). Most of the work on repeat instability in colorectal cancer focuses on variation between tumor and matched normal genomes in five marker repeats (Umar et al. 2004), which captures only a fraction of variation from more than 3 million human microsatellite loci (Payseur et al. 2011). To date, there are few studies that focus on genome-wide tandem repeat instability in cancers. One such study (McIver et al. 2014) compared repeat variation in breast cancer exomes with that in healthy tissues. Two other studies focused on repeats in colorectal cancer (Di Pietro et al. 2005; Woerner et al. 2003), but did not report repeat variation between tumor and healthy tissues.

Regulatory repeat variation in tumors has been poorly studied in the context of gene expression alterations. Recent advances in next generation sequencing and accurate repeat genotyping algorithms enabled us to investigate repeat variation in tumor genomes and their potential consequences on gene expression divergence. Here we analyze tandem repeats and repeat variation in 37 colorectal tumors and their matched normal genomes in the upstream regulatory regions of 18,709 genes, as well as in a smaller subset of genes in known cancer-associated pathways.

We found a significant enrichment for de novo repeat gain, repeat loss and copy number variation between tumor and their matched normal genomes compared to normal/normal genome pairs. We observed that genes with repeat instability are

overexpressed. Moreover, most well-studied cancer pathways, including the p53 and Wnt pathway are significantly enriched in repeat instability and show overexpression in those genes with repeat instability.

5.3. Methods

Genome sequence analysis

We obtained whole genome sequences of colon and rectal tumors, together with matched genomes -- the same individual's genomic sequences from blood samples -- from the controlled access data tier of the Cancer Genome Atlas Data Portal (TCGA, https://tcga-data.nci.nih.gov/tcga/). The genome sequence data is based on 3-5X coverage Illumina HiSeq2000 sequencing of 80-100 million 2X100 long base pair reads, aligned against human genome build #18 (Cancer and Atlas 2012) using the indel-compatible software package BWA (bwa-0.5.9rcl (Li 2012)).

We generated consensus sequences for gene promoters in the tumors and matched normal genomes using SAMtools (Li et al. 2009a). Because our previous work on human tandem repeats (Bilgin et al. 2014) suggests that the 5,000 base pairs [bps] upstream from the transcription start site contain most regulatory signals, we focused on this region, and refer to it as the promoter. While generating the consensus sequences, we noticed that some genomes contained many more unaligned sequences than others. We eliminated genomes, whose promoters contain unaligned nucleotides

that comprise more than 10 percent of the whole promoter, which reduced our data set to 37 genomes (see Supplementary Table 1, for a list of genomes). After removing genes from the data set whose promoter sequences could not be aligned, we focused our analysis on the remaining "global" set of 18,709 genes. Apart from analyzing this global set, we also performed a more detailed analysis of 375 cancer genes (Supplementary Table 2) that fall into five well-studied cancer associated pathways (Cancer and Atlas 2012; Fearon 2011; Vogelstein and Kinzler 2004).

Tandem Repeat Identification

We used the program Tandem Repeat Finder 4.07b (Gelfand et al. 2007) to identify tandem repeats in the consensus promoters. Specifically, we identified repeats with (i) Tandem Repeat Finder scores exceeding 80, (ii) an incidence of indels in adjacent repeat units below 10 percent (e.g., a repeat unit of 20 nucleotides can have up to two indels relative to the consensus pattern, which is the repeat unit most common in the whole repeat sequence (Gelfand et al. 2007)), and (iii) a sequence identity of repeat units above 90 percent (e.g., at least 18 nucleotides of a repeat unit of 20 nucleotides must match the consensus pattern). One motivation for these stringent thresholds is that we were most interested in how repeat variation might cause gene expression differences, and variation of tandem repeats increases strongly for repeats of high sequence similarity and Tandem Repeat Finder Scores (O'Dushlaine and Shields 2008). We considered both micro- and minisatellites with tandem repeat units up to 100 nucleotides in length. Longer repeats are more stable and therefore less likely to cause expression differences (O'Dushlaine and Shields 2008).

Repeat variation

To compute repeat copy number variation, we first identified repeats that had the same repeat units and that occurred upstream of the same genes in each tumor and its matched normal genome. We allowed positional variation of repeats up to 50 nucleotides within a promoter, because indels can cause substantial shifting even within a species (Durbin et al. 2010). We then computed the difference in repeat copy number between a tumor genome and its matched normal genome for each of these repeats. To this end, we computed for each gene its repeat variation between each genome pair, to generate an array of size 37. A 0 in those arrays indicates either there is no repeat in both gene pairs or the repeat has the same copy number in both genes. A value greater than 0 indicates repeat copy number variation between the gene pairs. For example, a gene that contains a sequence that is repeated three times in one genome and five times in another genome would have repeat variation of 2. Because none of the genes we analyzed had more than one repeat that varied in copy number, we were able to uniquely assign repeat variation values to genes.

Gene expression analysis

The gene expression data we used is based on RNA sequencing of 350-450 base pair-long Illumina Cluster Station and Genome Analyzer reads by The Cancer Genome Atlas (TCGA) Consortium (Cancer and Atlas 2012). The data comprises expression levels in reads per kilobase of transcript per million reads mapped (rpkm) for 18,709 genes in the 37 tumor genomes, whose genomes we analyzed.

5.4. Results

De novo repeat gain and loss are more frequent in tumor/normal genome pairs.

We identified genes with tandem repeats in the 5000 bp upstream from the transcription start site of n = 18,709 genes in 37 colorectal tumors and their matched normal genomes. We found that a tumor genome has on average 4192 promoters with tandem repeats, a number that is very similar to the 4165 promoters with tandem repeats in matched normal genomes.

A mean number of 1043 (±337 s.dev.) of genes in a tumor genome show repeats that do not occur in the same gene's promoter in the matched normal genome, compared to a mean number of 1016 (±334 s.dev.) promoter repeats that are specific to normal genomes and do not occur in tumor genomes. In total, there are 2059 (±373 s.dev.) genes which either lost a repeat or gained a de novo repeat in a tumor compared to their matched genes in normal genome. This number is significantly higher than the number of such genes with repeat losses or gains within normal genome pairs (1274 ±481 s.dev., Wilcoxon Rank Sum (WRS) test (Mann and Whitney 1947), $P < 10^{-16}$), when the 37 normal genomes are paired in all possible (666) combinations (see Figure 1).

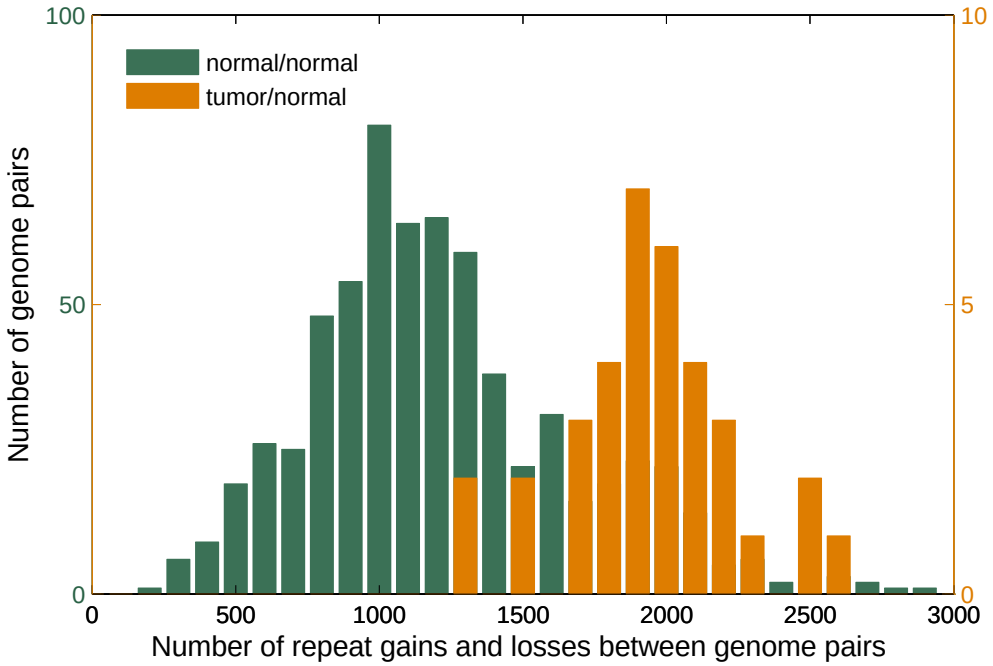

Figure 1. Repeat gain and losses are significantly more frequent between tumor and normal genomes. Histograms of the number of genes within a normal-normal pair (green bars, left axis, n = 666) and within a tumor-matched normal pair (orange bars, right axis, n = 37) with repeat gains and losses. A WRS test shows that the two distributions are significantly different ($P < 10^{-16}$).

Tumors are enriched for repeat copy number variation.

For those genes where both the tumor and matched normal genomes contain a repeat, we next asked how many repeats vary in the copy number of their repeat unit. Averaged over all 37 tumor/normal genome pairs, the number of genes with repeat variation is 157.6 (±23.7 s.dev.), or approximately 5 percent of all 2916 (±752 s.dev.) gene pairs where both members carry a repeat. This number is significantly higher than the number of genes with repeat variation between all possible pair combinations of normal genomes, which is 80.5 (±18.5 s.dev.) out of 2996.5 (±790 s.dev.) gene

pairs, that is ~ 2.7 percent (WRS test, $P < 10^{-24}$, see Figure 2). Overall, we conclude that tumor genomes harbor more repeat copy number variation than normal genomes.

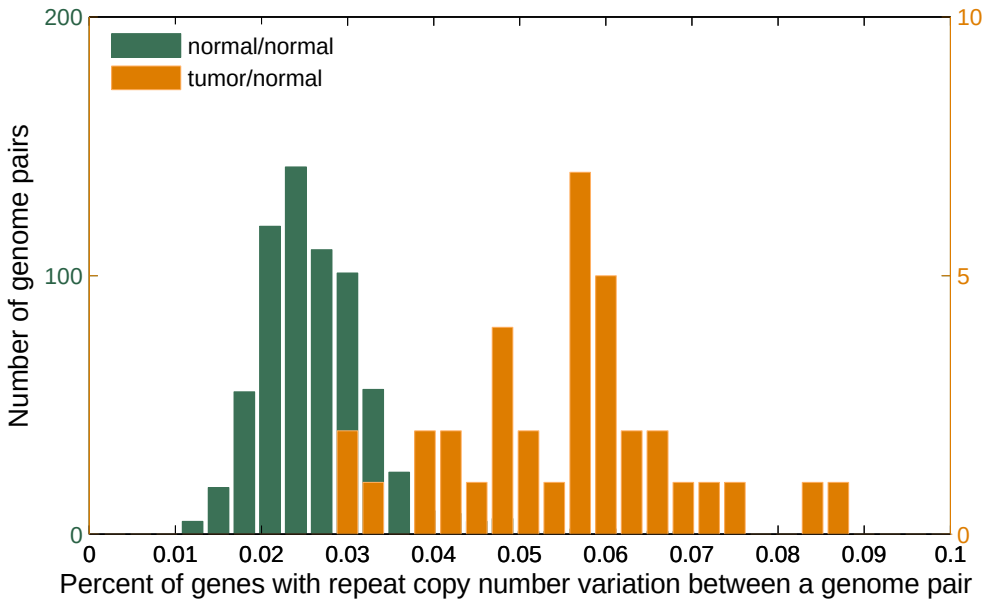

Figure 2. Repeat variation is significantly more frequent in tumor genomes. Histograms of proportion of genes within a normal-normal genome pair (green bars, left axis, n = 666) and within a tumor-matched normal pair (orange bars, right axis, n = 37) with repeat copy number variation. A WRS test shows that tumor genomes contain significantly more repeat variation ($P < 10^{-24}$).

Genes with repeat instability are significantly overexpressed.

Gene expression data sometimes provide tumor *signatures*, i.e. most tumors show expression patterns that are unique to their cancer type (Chung et al. 2002). Several studies reported gene expression changes due to tandem repeat mutations in gene promoters in healthy tissues of various organisms (Fondon et al. 2008; Gemayel et al. 2010; Vinces et al. 2009). We wondered, whether repeat instability in tumors has an effect on gene expression and in what direction they change gene expression. To this

end, we identified genes whose promoters contain any of several possible instability (de novo repeat gain, repeat loss or copy number variation) in at least one tumor/normal genome pair. We found 7258 such genes. Next, we retrieved RNA-seq based gene expression data from the Cancer Genome Atlas Data Portal (TCGA) (Cancer and Atlas 2012) for these genes in the 37 tumors. We computed for each gene, the binary logarithm of the mean expression level for those genes expressed in genomes, where the gene has a repeat instability and for the remaining, where the gene doesn't show any such instability. We then compared those mean gene expression levels with a Wilcoxon signed rank (WSR) test (Woolson 2008) and found that they differ significantly ($P < 10^{-104}$). Moreover, the genes with repeat instability were significantly overexpressed (see Figure 3).

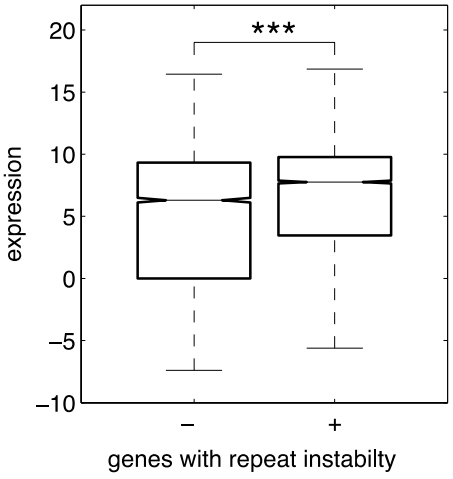

Figure 3. Genes with repeat instability are overexpressed.

Box plot of binary logarithm of mean expression levels of those genes expressed in genomes, where the gene has a repeat instability (right box) and for the remaining, where the gene doesn't show any such instability (left box) in at least one of the 37 tumor/normal genome pairs. Horizontal lines in the middle of each box mark the median, edges of boxes correspond to the 25th and 75th percentiles, and whiskers cover 99.3 percent of the data's range. *** indicates a highly significant difference ($P < 10^{-104}$, n = 7258) between the two data categories based on a WSR test.

Wnt signaling and p53 pathways show significantly higher repeat variation.

Many unique mutations in a cancer associated signaling pathway have the same functional effect on the pathway (Kan et al. 2010; Vogelstein and Kinzler 2004), which makes the analysis of entire pathways important to understand tumorigenesis (Cancer and Atlas 2012; Dhillon et al. 2007; Fearon 2011; Fresno Vara et al. 2004; Logan and Nusse 2004; Vogelstein and Kinzler 2004). We therefore next focused on five well-studied signaling pathways that are known to play a central role in carcinogenesis: Wnt, TGF beta, MAPK, mTOR, and p53 pathways, (Supplementary Table S2; (Cancer and Atlas 2012; Fearon 2011)). We identified 375 genes in these five pathways, which we will refer to as *cancer genes*. We found that, out of these 375 cancer genes, 126 show repeat instability (de novo repeat gain, repeat loss or copy number variation) between at least one tumor and one matched normal genome.

To find out whether particular cancer pathways are enriched for repeat variation, we calculated repeat variation between tumor/normal genome pairs, which is an array of size 37 for each cancer gene (see Methods). We then computed the mean of each array to arrive at a mean repeat variation value for each cancer gene. Similarly, we calculated the mean repeat variation of each cancer gene for 666 normal genome pairs. A comparison of these mean repeat variations of genes found in each pathway revealed that Wnt signaling and p53 pathways were significantly enriched for variable repeats (WSR test, P = 0.007 and 0.04 respectively, after Bonferroni correction, see Figure 4). TGF beta, on the other hand, showed no repeat variation either in tumor/normal pairs or in normal genome pairs. For MAPK and mTOR pathways,

mean repeat variation between tumor and matched normal genomes were both greater than mean repeat variation in normal genomes, but not significantly. These findings indicate an overall enrichment for variable repeats in most cancer-associated signaling pathways.

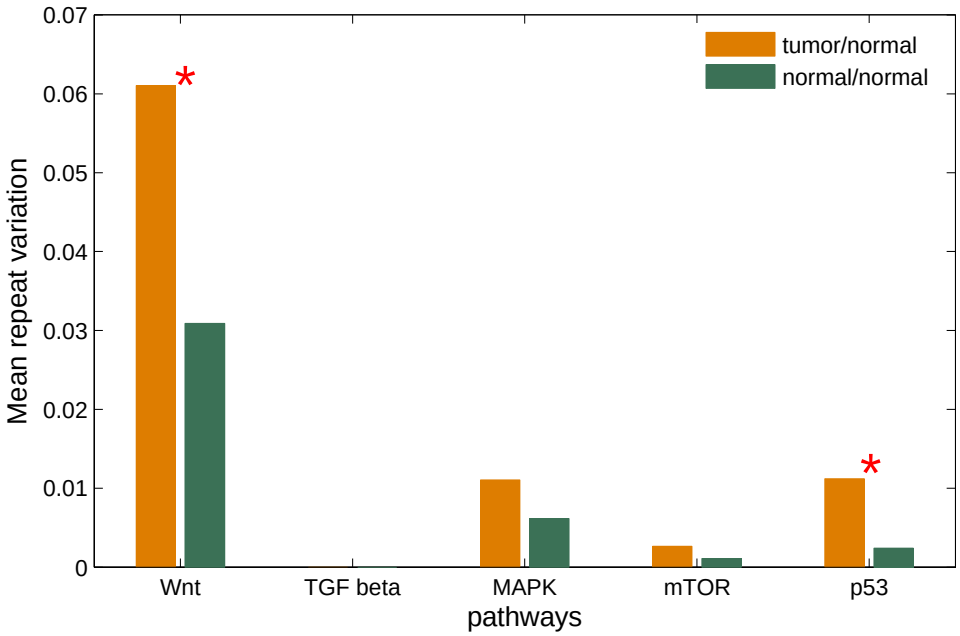

Figure 4. WNT and p53 pathways are significantly enriched for variable repeats. Bar plot of repeat variation for all 375 cancer genes in five cancer pathways averaged over all 37 tumor/normal genome pairs (orange bars) and 666 normal genome pairs (green bars). Red star indicates a significant difference based on a WSR test (after Bonferroni correction).

Four out of five cancer pathways are enriched for repeat gain and loss.

To find out whether particular cancer pathways are enriched for de novo repeat gain and repeat loss, we calculated first, for each cancer gene, in how many tumor/normal genome pairs it has a gain or loss. We then calculated in how many normal genome

pairs genes have a gain or loss. When we compared the proportion of genes with repeat gain or loss in each pathway, we found that four out of five pathways in tumor/normal genome pairs were enriched for repeat gain or loss compared to normal genome pairs. These are the Wnt pathway (WSR test, $P = 0.015$, after Bonferroni correction, see Figure 5), the MAPK pathway ($P = 0.01$), the mTOR pathway ($P < 10^{-7}$) and the p53 pathway ($P < 10^{-7}$). Only the TGF beta pathway did not show any significant difference.

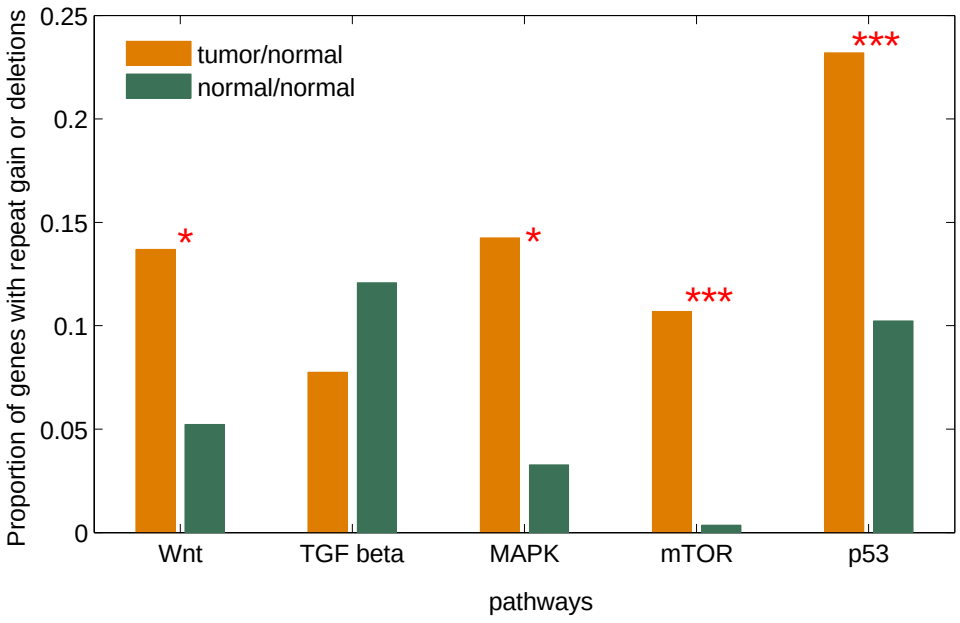

Figure 5. Most cancer pathways are significantly enriched for repeat repeat gain and loss. Bar plot of proportion of genes with repeat gain or loss for all 375 cancer genes in 37 tumor/normal genome pairs (orange bars) and in 666 normal genome pairs (green bars). * indicates a significant difference between two data categories based on a WSR test ($P < 0.05$), whereas *** indicates a highly significant difference ($P < 10^{-7}$ after Bonferroni correction).

Genes with repeat instability are significantly overexpressed in the Wnt, mTOR and p53 pathways.

In one of the above analyses, we showed that genes with repeat instability have significantly increased expression levels. We wondered whether this also holds for each individual cancer pathway. We therefore repeated our expression analysis for each pathway and for cancer genes with repeat instability. We found that in the Wnt (WSR test, $P = 0.03$, after Bonferroni correction), mTOR ($P < 10^{-5}$), and p53 ($P < 10^{-4}$) signaling pathways, genes with such instability were significantly overexpressed compared to the genes with no such instability, whereas the TGF beta ($P = 0.56$) and the MAPK ($P = 0.77$) pathways did not show a significant difference (see Figure 6).

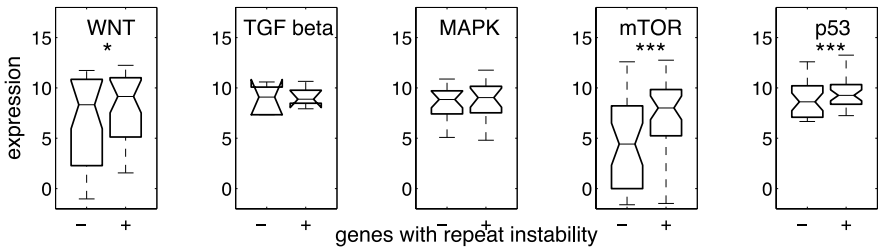

Figure 6. Genes with repeat gain, loss or variation are overexpressed in the Wnt, mTOR and p53 pathways. Box plot of binary logarithm of mean expression levels of cancer genes expressed in genomes, where the gene has a repeat instability (left box in each panel) and for the remaining, where the gene doesn't show any such instability (right box in each panel) in at least one of the 37 tumor/normal genome pairs for the Wnt (n=32, for both boxes), TGF beta (n=6), MAPK (n=19), mTOR (n=38) and p53 (n=31) signaling pathways. Horizontal lines in the middle of each box mark the median, edges of boxes correspond to the 25th and 75th percentiles, and whiskers cover 99.3 percent of the data's range. * indicates a significant difference between two data categories based on a WSR test ($P < 0.01$), whereas *** indicates a highly significant difference ($P < 10^{-4}$) after Bonferroni correction.

5.5. Discussion

We identified tandem repeat variation between colorectal tumor and healthy tissues in promoter regions of 18,709 human genes and their upstream regions, and in a smaller set of 375 genes and five signaling pathways associated with cancer. We found evidence that tumors are associated with greatly enhanced de novo evolution or loss of promoter repeats, as such events were significantly more abundant between tumor and matched normal genomes than between normal genome pairs. There were on average 2059 repeat gain or loss between tumor/normal genome pairs, whereas between normal genomes we found only 1274 gain or loss. We also observed a significant enrichment in repeat copy number variation in tumor/normal genome pairs. 5 percent of the repeats we identified vary in copy number between tumor and their matched normal genome, whereas only 2.7 percent of the repeats identified in normal genomes showed such copy number variation. Furthermore, genes are significantly overexpressed in tumor genomes where they show repeat instability compared to tumor genomes where they don't show any such instability.

Identification of mutated cancer genes provides insights into the biological processes underlying tumorigenesis (Futreal et al. 2004). However, the catalogue of mutated genes can be quite diverse and heterogenous even within same type of tumor (Jass 2007; Kan et al. 2010; Vogelstein and Kinzler 2004), whereas certain pathway dysregulations are shared among multiple cancer types (Kan et al. 2010; Van Limbergen et al. 2002; Segditsas and Tomlinson 2006; Vogelstein and Kinzler 2004). We therefore analyzed five cancer-associated pathways for repeat instability in the promoters of pathway-associated genes. One of them is the Wnt signaling pathway,

which is commonly implicated in carcinogenesis due to its regulatory role in cell proliferation, gene transcription and cell migration (Cancer and Atlas 2012; Logan and Nusse 2004). Colorectal cancers of all subtypes almost invariably start with an activating mutation in this pathway (Fearon 2011; Sadanandam et al. 2013), causing dysregulation of the pathway (Logan and Nusse 2004). Remarkably, we found that genes in the Wnt pathway are significantly enriched for repeat instability and genes in this pathway with repeat instability are significantly overexpressed. The MAPK (Kan et al. 2010) and mTOR (Fresno Vara et al. 2004) pathways, which are often hyperactivated in cancer cells, show higher repeat instability in tumor/normal genome pairs than in normal genome pairs. Genes with repeat instability are significantly overexpressed in the mTOR pathway. Conversely, none of the genes in the TGF beta pathway show increased repeat instability or expression alterations that associate with repeat instability. This observation is in line with a previous finding (Cancer and Atlas 2012), which suggests that this pathway is the least divergent pathway between colorectal tumors and their matched normal genomes in terms of copy number variation and gene expression. The final pathway we analyzed, (p53) plays a crucial role in the cell cycle and can initiate cell death (Harris and Levine 2005; Vazquez et al. 2008). Inactivation of p53 pathway through multiple mutations is an almost universal feature of human cancer cells (Grochola et al. 2010; Whibley et al. 2009). In agreement with this, we found that genes in the p53 pathway are significantly enriched for unstable repeats in tumor/normal pairs compared to normal genome pairs and genes with such repeats are significantly overexpressed.

Overexpression through molecular alterations is a common phenomenon in carcinogenesis. For example, in the breast cancer *Rac1* gene, a molecular switch to

control cell growth is activated through an insertion of 19 codons to its open reading frame, which is implicated in disease aggression (Schnelzer et al. 2000). Another example comes from lung cancer, where small deletions or substitutions clustered around ATP-binding pocket of *EGFR*, a *receptor tyrosine kinase* gene, activates the tyrosine kinase activity, leading to increased growth factor signaling (Lynch et al. 2004). Likewise, in basal cell carcinoma, activating mutations in the *Smoothened* gene cause overexpression, which then functions as an oncogene (Xie et al. 1998).

Among the limitations of our study is that we cannot distinguish between somatic and germline mutations. This is relevant, because some mismatch repair genes can experience germline mutations that cause colorectal cancer (Vilar and Gruber 2010). These germline mutations also play a role in forming different subtypes of colorectal cancer, as they trigger accumulation of different sets of somatic mutations throughout carcinogenesis (Fearon 2011). However, because 90 percent of cancer mutations are somatic (Futreal et al. 2004), this should not be a serious limitation. Second, our cancer gene set is unlikely to encompass all genes that may play a role in cancer, because we focused on particular, well studied cancer associated pathways. Another obstacle of our study is the limited number of genomes we could analyze, and the lack of gene expression information in matched normal genomes. Finally, limitations in genome alignment quality may cause false detection of repeat copy numbers and this may increase apparent repeat variation. Information coming from more and higher quality genomes will enable more precise identification of repeat instability and allow researchers to associate them better with expression variation in a causative manner.

As genetic instability is not only central to pathogenesis but also may underlie the development of resistance to chemotherapeutic agents, identification of mutational mechanisms responsible for it is an important area of study. We believe that our findings will increase the understanding of the molecular basis of genetic instability in carcinogenesis, and thereby facilitate the development of more precise and effective molecular diagnostic and therapeutic approaches.

5.6. References

Alqurashi, N., Gopalan, V., Smith, R. A., & Lam, A. K. Y. (2013). Clinical impacts of mammalian target of rapamycin expression in human colorectal cancers. *Human Pathology*, *44*, 2089–2096. doi:10.1016/j.humpath.2013.03.014

Bilgin, T., Robinson, M. D., & Wagner, A. (2014). Tandem repeats and increased expression divergence in primate genes.

Burgess, D. J. (2013). Gene expression: colorectal cancer classifications. *Nature Reviews. Cancer*, *13*, 380–1. doi:10.1038/nrc3529

Cancer, T., & Atlas, G. (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, *487*, 330–7. doi:10.1038/nature11252

Chung, C. H., Bernard, P. S., & Perou, C. M. (2002). Molecular portraits and the family tree of cancer. *Nature Genetics*, *32 Suppl*, 533–540. doi:10.1038/ng1038

Dhillon, A. S., Hagan, S., Rath, O., & Kolch, W. (2007). MAP kinase signalling pathways in cancer. *Oncogene*, *26*, 3279–3290. doi:10.1038/sj.onc.1210421

Di Pietro, M., Bellver, J. S., Menigatti, M., Bannwart, F., Schnider, A., Russell, A., … Marra, G. (2005). Defective DNA mismatch repair determines a characteristic transcriptional profile in proximal colon cancers. *Gastroenterology*, *129*, 1047–1059. doi:10.1053/j.gastro.2005.06.028

Durbin, R. M., Altshuler, D. L., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., … Peterson, J. L. (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*(7319), 1061–1073. Retrieved from http://www.nature.com/doifinder/10.1038/nature09534

Fearon, E. R. (2011). Molecular genetics of colorectal cancer. *Annual Review of Pathology*, *6*, 479–507. doi:10.1146/annurev-pathol-011110-130235

Fondon, J. W., Hammock, E. a D., Hannan, A. J., & King, D. G. (2008). Simple sequence repeats: genetic modulators of brain function and behavior. *Trends in Neurosciences*, *31*(7), 328–34. doi:10.1016/j.tins.2008.03.006

Fresno Vara, J. A., Casado, E., de Castro, J., Cejas, P., Belda-Iniesta, C., & González-Barón, M. (2004). PI3K/Akt signalling pathway and cancer. *Cancer Treatment Reviews*, *30*, 193–204. doi:10.1016/j.ctrv.2003.07.007

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., … Stratton, M. R. (2004). A census of human cancer genes. *Nature Reviews. Cancer*, *4*, 177–183. doi:10.1038/nrc1299

Gelfand, Y., Rodriguez, A., & Benson, G. (2007). TRDB—The Tandem Repeats Database. *Nucleic Acids Research*, *35*(Database issue), D80–D87. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17175540

Gemayel, R., Vinces, M. D., Legendre, M., & Verstrepen, K. J. (2010). Variable Tandem Repeats Accelerate Evolution of Coding and Regulatory Sequences. *Annual Review of Genetics*. doi:10.1146/annurev-genet-072610-155046

Giovannucci, E., Stampfer, M. J., Krithivas, K., Brown, M., Dahl, D., Brufsky, A., … Kantoff, P. W. (1997). The CAG repeat within the androgen receptor gene and its relationship to prostate cancer. *Proceedings of the National Academy of Sciences of the United States of America*, *94*, 3320–3323. doi:10.1073/pnas.94.7.3320

Grochola, L. F., Zeron-Medina, J., Mériaux, S., & Bond, G. L. (2010). Single-nucleotide polymorphisms in the p53 signaling pathway. *Cold Spring Harbor Perspectives in Biology*, *2*, a001032. doi:10.1101/cshperspect.a001032

Harris, S. L., & Levine, A. J. (2005). The p53 pathway: positive and negative feedback loops. *Oncogene*, *24*, 2899–2908. doi:10.1038/sj.onc.1208615

Hewish, M., Lord, C. J., Martin, S. A., Cunningham, D., & Ashworth, A. (2010). Mismatch repair deficient colorectal cancer in the era of personalized treatment. *Nature Reviews. Clinical Oncology*, *7*, 197–208. doi:10.1038/nrclinonc.2010.18

Imai, K., & Yamamoto, H. (2008). Carcinogenesis and microsatellite instability: The interrelationship between genetics and epigenetics. *Carcinogenesis*. doi:10.1093/carcin/bgm228

Jass, J. R. (2007). Classification of colorectal cancer based on correlation of clinical, morphological and molecular features. *Histopathology*. doi:10.1111/j.1365-2559.2006.02549.x

Jorissen, R. N., Lipton, L., Gibbs, P., Chapman, M., Desai, J., Jones, I. T., … Sieber, O. M. (2008). DNA copy-number alterations underlie gene expression differences between microsatellite stable and unstable colorectal cancers. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*, *14*, 8061–8069. doi:10.1158/1078-0432.CCR-08-1431

Kan, Z., Jaiswal, B. S., Stinson, J., Janakiraman, V., Bhatt, D., Stern, H. M., … Seshagiri, S. (2010). Diverse somatic mutation patterns and pathway alterations in human cancers. *Nature*, *466*, 869–873. doi:10.1186/gb-2010-11-s1-p37

Krontiris, T. G., Devlin, B., Karp, D. D., Robert, N. J., & Risch, N. (1993). An association between the risk of cancer and mutations in the HRAS1 minisatellite locus. *The New England Journal of Medicine*, *329*(8), 517–523. Retrieved from http://www.nejm.org/doi/full/10.1056/NEJM199308193290801

Laplante, M., & Sabatini, D. M. (2012). MTOR signaling in growth control and disease. *Cell*. doi:10.1016/j.cell.2012.03.017

Legendre, M., Pochet, N., Pak, T., & Verstrepen, K. J. (2007). Sequence-based estimation of minisatellite and microsatellite repeat variability. *Genome Research*, *17*(12), 1787–1796. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2099588&tool=pmcentrez&rendertype=abstract

Li, H. (2012). Exploring single-sample snp and indel calling with whole-genome de novo assembly. *Bioinformatics*, *28*, 1838–1844. doi:10.1093/bioinformatics/bts280

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment / Map (SAM) Format and SAMtools 1000 Genome Project Data Processing Subgroup. *Bioinformatics*, *25*, 2078–2079. doi:10.1093/bioinformatics/btp352

Li, Y.-C., Korol, A. B., Fahima, T., Beiles, A., & Nevo, E. (2002). Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Molecular Ecology*, *11*(12), 2453–65. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12453231

Logan, C. Y., & Nusse, R. (2004). The Wnt signaling pathway in development and disease. *Annual Review of Cell and Developmental Biology*, *20*, 781–810. doi:10.1146/annurev.cellbio.20.010403.113126

López Castel, A., Cleary, J. D., & Pearson, C. E. (2010). Repeat instability as the basis for human diseases and as a potential target for therapy. *Nature Reviews Molecular Cell Biology*, *11*(3), 165–170. Retrieved from http://dx.doi.org/10.1038/nrm2854

Lynch, T. J., Bell, D. W., Sordella, R., Gurubhagavatula, S., Okimoto, R. A., Brannigan, B. W., … Haber, D. A. (2004). Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *The New England Journal of Medicine*, *350*, 2129–2139. doi:10.1056/NEJMoa040938

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, *18*(1), 50–60. doi:10.1214/aoms/1177730491

McIver, L. J., Fonville, N. C., Karunasena, E., & Garner, H. R. (2014). Microsatellite genotyping reveals a signature in breast cancer exomes. *Breast Cancer Research and Treatment*, *145*(3), 791–8. doi:10.1007/s10549-014-2908-8

Nosho, K., Yamamoto, H., Adachi, Y., Endo, T., Hinoda, Y., & Imai, K. (2005). Gene expression profiling of colorectal adenomas and early invasive carcinomas by cDNA array analysis. *British Journal of Cancer*, *92*, 1193–1200. doi:10.1038/sj.bjc.6602442

O'Dushlaine, C. T., & Shields, D. C. (2008). Marked variation in predicted and observed variability of tandem repeat loci across the human genome. *BMC Genomics*, *9*, 175. doi:10.1186/1471-2164-9-175

Payseur, B. a, Jing, P., & Haasl, R. J. (2011). A genomic portrait of human microsatellite variation. *Molecular Biology and Evolution*, *28*(1), 303–12. doi:10.1093/molbev/msq198

Sadanandam, A., Lyssiotis, C. A., Homicsko, K., Collisson, E. A., Gibb, W. J., Wullschleger, S., … Hanahan, D. (2013). A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nature Medicine*, *19*, 619–25. doi:10.1038/nm.3175

Schlötterer, C. (2000). Evolutionary dynamics of microsatellite DNA. *Chromosoma*, *109*(6), 365–371. doi:10.1007/s004120000089

Schnelzer, A., Prechtel, D., Knaus, U., Dehne, K., Gerhard, M., Graeff, H., … Lengyel, E. (2000). Rac1 in human breast cancer: overexpression, mutation analysis, and characterization of a new isoform, Rac1b. *Oncogene*, *19*, 3013–3020. doi:10.1038/sj.onc.1203621

Segditsas, S., & Tomlinson, I. (2006). Colorectal cancer and genetic alterations in the Wnt pathway. *Oncogene*, *25*, 7531–7537. doi:10.1038/sj.onc.1210059

Siegel, R., Naishadham, D., & Jemal, A. (2013). Cancer statistics, 2013. *CA: A Cancer Journal for Clinicians*, *63*, 11–30. doi:10.3322/caac.21166

Tian, S., Roepman, P., Popovici, V., Michaut, M., Majewski, I., Salazar, R., … Simon, I. (2012). A robust genomic signature for the detection of colorectal cancer patients with microsatellite instability phenotype and high mutation frequency. *Journal of Pathology*, *228*, 586–595. doi:10.1002/path.4092

UK, C. R. (2014). Worldwide cancer statistics. *Cancer research UK*.

Umar, A., Boland, C. R., Terdiman, J. P., Syngal, S., de la Chapelle, A., Rüschoff, J., … Srivastava, S. (2004). Revised Bethesda Guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability. *Journal of the National Cancer Institute*. doi:10.1093/jnci/djh034

Van Limbergen, H., Poppe, B., Michaux, L., Herens, C., Brown, J., Noens, L., … Speleman, F. (2002). Frequent alterations in the Wnt signaling pathway in colorectal cancer with microsatellite instability. *Genes Chromosomes and Cancer*, *33*, 73–81. doi:10.1002/gcc.1226

Vazquez, A., Bond, E. E., Levine, A. J., & Bond, G. L. (2008). The genetics of the p53 pathway, apoptosis and cancer therapy. *Nature Reviews. Drug Discovery*, *7*, 979–987. doi:10.1038/nrd2656

Vilar, E., & Gruber, S. B. (2010). Microsatellite instability in colorectal cancer-the stable evidence. *Nature Reviews. Clinical Oncology*, *7*, 153–162. doi:10.1038/nrclinonc.2009.237

Vinces, M. D., Legendre, M., Caldara, M., Hagihara, M., & Verstrepen, K. J. (2009). Unstable tandem repeats in promoters confer transcriptional evolvability. *Science (New York, N.Y.)*, *324*(5931), 1213–6. doi:10.1126/science.1170097

Vogelstein, B., & Kinzler, K. W. (2004). Cancer genes and the pathways they control. *Nature Medicine*, *10*, 789–799. doi:10.1038/nm1087

Wang, X.-W., & Zhang, Y.-J. (2014). Targeting mTOR network in colorectal cancer therapy. *World Journal of Gastroenterology : WJG*, *20*, 4178–4188. doi:10.3748/wjg.v20.i15.4178

Whibley, C., Pharoah, P. D. P., & Hollstein, M. (2009). p53 polymorphisms: cancer implications. *Nature Reviews. Cancer*, *9*, 95–107. doi:10.1038/nrc2584

Woerner, S. M., Benner, A., Sutter, C., Schiller, M., Yuan, Y. P., Keller, G., … Gebert, J. F. (2003). Pathogenesis of DNA repair-deficient cancers: a statistical meta-analysis of putative Real Common Target genes. *Oncogene*, *22*, 2226–2235. doi:10.1038/sj.onc.1206421

Woolson, R. F. (2008). Wilcoxon signed-rank test. *Wiley Encyclopedia of Clinical Trials*, 1–3. doi:10.1002/9780471462422.eoct979

Xie, J., Murone, M., Luoh, S. M., Ryan, A., Gu, Q., Zhang, C., … de Sauvage, F. J. (1998). Activating Smoothened mutations in sporadic basal-cell carcinoma. *Nature*, *391*, 90–92. doi:10.1038/34201

Zitt, M. M., Untergasser, G., Amberger, A., Moser, P., Stadlmann, S., Müller, H. M., … Ofner, D. (2008). Dickkopf-3 as a new potential marker for neoangiogenesis in colorectal cancer: expression in cancer tissue and adjacent non-cancerous tissue. *Disease Markers*, *24*, 101–109.

5.7. Supplementary Material

Genome	cancer type	Genome	cancer type
A6_2676	colon	AF_2691	rectum
A6_2678	colon	AF_2692	rectum
A6_3807	colon	AG_3574	rectum
AA_3514	colon	AG_3728	rectum
AA_3516	colon	AG_3878	rectum
AA_3529	colon	AG_3887	rectum
AA_3548	colon	AG_3892	rectum
AA_3549	colon	AG_3894	rectum
AA_3555	colon	AG_3909	rectum
AA_3664	colon	AG_4001	rectum
AA_3666	colon	AG_4005	rectum
AA_3861	colon	AG_4007	rectum
AA_3947	colon	AG_4008	rectum

AA_3956	colon	AG_4015	rectum
AA_3968	colon	AG_A002	rectum
AA_A00U	colon	AG_A00Y	rectum
AA_A00Z	colon	AG_A011	rectum
AA_A01K	colon	AG_A032	rectum
AA_A02R	colon		

Supplementary Table S1. List of genomes considered in the study. The left column indicates TCGA sequence IDs, the right column cancer type (colon or rectal).

pathway	gene	pathway	gene	pathway	gene
Wnt	APC2	TGF beta	TGFA	MAPK	JUN
Wnt	APCDD1L	TGF beta	TGFB1I1	MAPK	PDCD4
Wnt	APCDD1	TGF beta	TGFB1	MAPK	MAPK10
Wnt	APCS	TGF beta	TGFB2	MAPK	MAPK11
Wnt	APC	TGF beta	TGFB3	MAPK	MAPK12
Wnt	WNT10A	TGF beta	TGFBI	MAPK	MAPK13
Wnt	WNT10B	TGF beta	TGFBR1	MAPK	MAPK14
Wnt	WNT11	TGF beta	TGFBR2	MAPK	MAPK15
Wnt	WNT16	TGF beta	TGFBR3	MAPK	MAPK1
Wnt	WNT1	TGF beta	TGFBRAP1	MAPK	MAPK3
Wnt	WNT2B	TGF beta	SMAD1	MAPK	MAPK4
Wnt	WNT2	TGF beta	SMAD2	MAPK	MAPK6
Wnt	WNT3A	TGF beta	SMAD3	MAPK	MAPK7
Wnt	WNT3	TGF beta	SMAD4	MAPK	MAPK8
Wnt	WNT4	TGF beta	SMAD5	MAPK	MAPK9
Wnt	WNT5A	TGF beta	SMAD6	MAPK	MAPKAP1
Wnt	WNT5B	TGF beta	SMAD7	MAPK	MAPKBP1
Wnt	WNT6	TGF beta	SMAD9	MAPK	MAP2K1
Wnt	WNT7A	TGF beta	TNFRSF1A	MAPK	MAP2K2
Wnt	WNT7B	TGF beta	TNFRSF1B	MAPK	MAP2K3
Wnt	WNT8A	TGF beta	TNF	MAPK	MAP2K4
Wnt	WNT8B	TGF beta	ACVR1B	MAPK	MAP2K5
Wnt	WNT9A	TGF beta	ACVR1C	MAPK	MAP2K6
Wnt	WNT9B	TGF beta	ACVR1	MAPK	MAP2K7
Wnt	CTNNB1	TGF beta	ACVR2A	MAPK	MAP3K10
Wnt	CTNNBL1	TGF beta	ACVR2B	MAPK	MAP3K11
Wnt	AXIN1	TGF beta	ACVRL1	MAPK	MAP3K12
Wnt	AXIN2	TGF beta	BMPR1A	MAPK	MAP3K13
Wnt	GSK3A	TGF beta	BMPR1B	MAPK	MAP3K14
Wnt	GSK3B	TGF beta	HRAS	MAPK	MAP3K15
Wnt	BTRC	p53	TP53AIP1	MAPK	MAP3K1
Wnt	CSNK1A1L	p53	TP53BP1	MAPK	MAP3K2
Wnt	CSNK1A1	p53	TP53BP2	MAPK	MAP3K3
Wnt	CSNK1D	p53	TP53I11	MAPK	MAP3K4

Wnt	CSNK1E	p53	TP53I13	MAPK	MAP3K5
Wnt	CSNK1G1	p53	TP53I3	MAPK	MAP3K6
Wnt	CSNK1G2	p53	TP53INP1	MAPK	MAP3K7
Wnt	CSNK1G3	p53	TP53INP2	MAPK	MAP3K8
Wnt	DVL1	p53	TP53RK	MAPK	MAP3K9
Wnt	DVL2	p53	TP53TG1	MAPK	MAP4K1
Wnt	DVL3	p53	TP53TG3B	MAPK	MAP4K2
Wnt	TCF12	p53	TP53TG5	MAPK	MAP4K3
Wnt	TCF15	p53	TP53	MAPK	MAP4K4
Wnt	TCF19	p53	MDM2	MAPK	MAP4K5
Wnt	TCF20	p53	ATMIN	MAPK	MAP4
Wnt	TCF21	p53	ATM	MAPK	MAP6D1
Wnt	TCF23	p53	CASP10	MAPK	MAP6
Wnt	TCF25	p53	CASP12	MAPK	MAP7D1
Wnt	TCF3	p53	CASP14	MAPK	MAP7D2
Wnt	TCF4	p53	CASP1	MAPK	MAP7D3
Wnt	TCF7L1	p53	CASP2	MAPK	MAP7
Wnt	TCF7L2	p53	CASP3	MAPK	MAP9
Wnt	TCF7	p53	CASP4	MAPK	KRAS
Wnt	TCFL5	p53	CASP5	MAPK	BRAF
Wnt	TLE1	p53	CASP6	MAPK	NRAS
Wnt	TLE2	p53	CASP7	MAPK	EGFR
Wnt	TLE3	p53	CASP8	MAPK	ERBB2
Wnt	TLE4	p53	CASP9	MAPK	ERBB3
Wnt	TLE6	p53	FASLG	MAPK	ERBB4
Wnt	CREBBP	p53	FASN	MAPK	FGF10
Wnt	EP300	p53	FASTKD1	MAPK	FGF11
Wnt	LRP10	p53	FASTKD2	MAPK	FGF12
Wnt	LRP11	p53	FASTKD3	MAPK	FGF13
Wnt	LRP12	p53	FASTKD5	MAPK	FGF14
Wnt	LRP1B	p53	FASTK	MAPK	FGF16
Wnt	LRP1	p53	FAS	MAPK	FGF17
Wnt	LRP4	p53	CDC20B	MAPK	FGF18
Wnt	LRP5L	p53	CDC20	MAPK	FGF19
Wnt	LRP5	p53	CDC23	MAPK	FGF1
Wnt	LRP6	p53	CDC25A	MAPK	FGF20
Wnt	LEF1	p53	CDC25B	MAPK	FGF21
Wnt	MT1B	p53	CDC25C	MAPK	FGF22
Wnt	NKD1	p53	CDC26	MAPK	FGF23
Wnt	NKD2	p53	CDC27	MAPK	FGF2
Wnt	DKK1	p53	BAX	MAPK	FGF3
Wnt	DKK2	p53	NOXA1	MAPK	FGF4
Wnt	DKK3	p53	BBC3	MAPK	FGF5
Wnt	DKK4	p53	CHEK1	MAPK	FGF6
Wnt	CTBP1	p53	CHEK2	MAPK	FGF7
Wnt	CTBP2	p53	SIRT1	MAPK	FGF8

Wnt	SFRP1	p53	CDK10	MAPK	FGF9
Wnt	SFRP2	p53	CDK11A	MAPK	FGFR1
Wnt	SFRP4	p53	CDK11B	MAPK	FGFR2
Wnt	SFRP5	p53	CDK12	MAPK	FGFR3
Wnt	RHOA	p53	CDK13	MAPK	FGFRL1
Wnt	RTKN2	p53	CDK14	MAPK	MYC
Wnt	RTKN	p53	CDK15	MAPK	RAF1
Wnt	CDX2	p53	CDK16	MAPK	RASA1
Wnt	FBXW2	p53	CDK17	MAPK	RASA2
mTOR	PIP4K2A	p53	CDK18	MAPK	RASA3
mTOR	PIP4K2B	p53	CDK19	MAPK	RASA4
mTOR	PIP4K2C	p53	CDK1	MAPK	RASD1
mTOR	PIP5K1A	p53	CDK20	MAPK	RASD2
mTOR	PIP5K1B	p53	CDK2	MAPK	RASEF
mTOR	PIP5K1C	p53	CDK3	MAPK	RASGEF1A
mTOR	PIP5K1P1	p53	CDK4	MAPK	RASGEF1B
mTOR	PIP5KL1	p53	CDK5	MAPK	RASGEF1C
mTOR	PIPOX	p53	CDK6	MAPK	RASGRF1
mTOR	PIPSL	p53	CDK7	MAPK	RASGRF2
mTOR	PIP	p53	CDK8	MAPK	RASGRP1
mTOR	PTENP1	p53	CDK9	MAPK	RASGRP2
mTOR	PTEN	p53	CDKL1	MAPK	RASGRP3
mTOR	MTOR	p53	CDKL2	MAPK	RASGRP4
mTOR	IGF1R	p53	CDKL3	MAPK	PRKAA1
mTOR	IGF1	p53	CDKL4	MAPK	PRKAA2
mTOR	IGF2R	p53	CDKL5	MAPK	PRKAB1
mTOR	IGF2	p53	CDKN1A	MAPK	PRKAB2
mTOR	IRS1	p53	CDKN1B	MAPK	PRKACA
mTOR	IRS2	p53	CDKN1C	MAPK	PRKACB
mTOR	IRS4	p53	CDKN2A	MAPK	PRKACG
mTOR	PIK3AP1	p53	CDKN2B	MAPK	PRKAG1
mTOR	PIK3C2A	p53	CDKN2C	MAPK	PRKAG2
mTOR	PIK3C2B	p53	CDKN2D	MAPK	PRKAG3
mTOR	PIK3C2G	p53	CDKN3		
mTOR	PIK3C3	p53	BCL2A1		
mTOR	PIK3CA	p53	BCL2L10		
mTOR	PIK3CB	p53	BCL2L11		
mTOR	PIK3CD	p53	BCL2L12		
mTOR	PIK3CG	p53	BCL2L13		
mTOR	PIK3R1	p53	BCL2L14		
mTOR	PIK3R2	p53	BCL2L15		
mTOR	PIK3R3	p53	BCL2L1		
mTOR	PIK3R4	p53	BCL2L2		
mTOR	PIK3R5	p53	BCL2		
mTOR	PIK3R6	p53	CCNE1		
mTOR	PDK1	p53	CCNE2		

mTOR	PDK2	p53	CCND1
mTOR	PDK3	p53	CCND2
mTOR	PDK4	p53	CCND3
mTOR	AKT1		
mTOR	AKT2		
mTOR	AKT3		
mTOR	STK11		

Supplementary Table S2. List of genes involved in cancer pathways, indicated with their HUGO (Eyre et al. 2006) gene IDs and the pathways they are associated with.